

---

# Probabilities on Sentences in an Expressive Logic

---

**Marcus Hutter**

Research School of Computer Science  
The Australian National University  
marcus.hutter@anu.edu.au

**John W. Lloyd**

Research School of Computer Science  
The Australian National University  
john.lloyd@anu.edu.au

**Kee Siong Ng**

EMC Greenplum and  
The Australian National University  
keesiong.ng@emc.com

**William T. B. Uther**

National ICT Australia and  
University of New South Wales  
william.uth@nicta.com.au

12 September 2012

## Abstract

<sup>1</sup> Automated reasoning about uncertain knowledge has many applications. One difficulty when developing such systems is the lack of a completely satisfactory integration of logic and probability. We address this problem directly. Expressive languages like higher-order logic are ideally suited for representing and reasoning about structured knowledge. Uncertain knowledge can be modeled by using graded probabilities rather than binary truth-values. The main technical problem studied in this paper is the following: Given a set of sentences, each having some probability of being true, what probability should be ascribed to other (query) sentences? A natural wish-list, among others, is that the probability distribution (i) is consistent with the knowledge base, (ii) allows for a consistent inference procedure and in particular (iii) reduces to deductive logic in the limit of probabilities being 0 and 1, (iv) allows (Bayesian) inductive reasoning and (v) learning in the limit and in particular (vi) allows confirmation of universally quantified hypotheses/sentences. We translate this wish-list into technical requirements for a prior probability and show that probabilities satisfying all our criteria exist. We also give explicit constructions and several general characterizations of probabilities that satisfy some or all of the criteria and various (counter) examples. We also derive necessary and sufficient conditions for extending beliefs about finitely many sentences to suitable probabilities over all sentences, and in particular least dogmatic or least biased ones. We conclude with a brief outlook on how the developed theory might be used and approximated in autonomous reasoning agents. Our theory is a step towards a globally consistent and empirically satisfactory unification of probability and logic.

---

<sup>1</sup>Presented at the Fifth Workshop on Combining Probability and Logic (Prolog 2011) in New York.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Logic</b>	<b>4</b>
<b>3</b>	<b>Probabilities on Sentences</b>	<b>8</b>
<b>4</b>	<b>Probabilities on Interpretations</b>	<b>16</b>
<b>5</b>	<b>Existence of Probabilities</b>	<b>22</b>
<b>6</b>	<b>Relative Entropy of Probabilities on Sentences</b>	<b>30</b>
<b>7</b>	<b>Extension of Probabilities</b>	<b>36</b>
<b>8</b>	<b>User Manual</b>	<b>41</b>
<b>9</b>	<b>Discussion</b>	<b>45</b>
	<b>References</b>	<b>47</b>
<b>A</b>	<b>List of Notation</b>	<b>50</b>
<b>B</b>	<b>List of Definitions, Theorems, Examples, ...</b>	<b>50</b>

## Keywords

higher-order logic; probability on sentences; Gaifman; Cournot; Bayes; induction; confirmation; learning; prior; knowledge; entropy.

*“The study of probability functions defined over the sentences of a rich enough formal language yields interesting insights in more than one direction.”*

— Haim Gaifman (1982)

## 1 Introduction

**Motivation.** Sophisticated computer applications generally require expressive languages for knowledge representation and reasoning. In particular, such languages need to be able to represent both structured knowledge and uncertainty [Nil86, Hal03, Mug96, DK03, RD06, Háj01, Wil02]. A suitable language for this purpose is higher-order logic [Chu40, Hen50, And02, Llo03, vBD83, Lei94, Sha01], which admits higher-order functions that can take functions as arguments and/or return functions as results. This facility is convenient for probabilistic modeling since it means that theories can contain probability densities [Far08, Pfe07, GMR<sup>+</sup>08]. In particular, many forms of probabilistic reasoning can be done in higher-order logic using the traditional axiomatic method: a theory can be written down which has the intended interpretation as a model and then conventional proof and computation techniques can be used to answer queries [NL09, NLU08]. While such a computational approach is effective, it is sometimes more natural to pose a problem as one where the probability of some sentences in the theory being true may be strictly less than one and/or the query sentence (and its negation) may not be a logical consequence of the theory. In such cases, deductive reasoning does not suffice for answering queries and it becomes necessary to use probabilistic methods [Par94, KD07, RD06, Mug96, MR07].

**Main aim.** These considerations lead to the main technical issue studied in this paper:

Given a set of sentences, each having some probability of being true,  
what probability should be ascribed to other (query) sentences?

We build on the work of Gaifman [Gai64] whose paper with Snir [GS82] develops a quite comprehensive theory of probabilities on sentences in first-order Peano arithmetic. We take up these ideas, using non-dogmatic priors [GS82] and additionally the minimum relative entropy

principle as in [Wil08a], but for general theories and in a higher-order setting. We concentrate on developing probabilities on sentences in a higher-order logic. This sets the stage for combining it with the probabilities inside sentences approach [NL09, NLU08].

**Summary of key concepts.** Section 2 introduces higher-order logic and its relevant properties. We use the higher-order logic (Definitions 1, 2, and 8) based on Church’s simple theory of types [Chu40, Hen50, And02]. We employ the Henkin semantics and make use of a particular class of interpretations, called separating interpretations (Definition 12).

Section 3 gives the definition of probabilities on sentences in higher-order logic (Definition 17), introduces the Gaifman condition, and develops some basic properties of such probabilities. Section 4 then introduces probabilities on interpretations and shows their close connection with probabilities on sentences. Gaifman [Gai64] (generalized in Definition 20 and Propositions 21, 22, 23) introduced a condition, called Gaifman in [SK66], that connects probabilities of quantified sentences to limits of probabilities of finite conjunctions. In our case, it effectively restricts probabilities to separating interpretations while maintaining countable additivity.

While generally accepted in probability theory (Definition 28), some circles argue that countable additivity (CA) does not have a good philosophical justification, and/or that it is not needed since real experience is always finite, hence only non-asymptotic statements are of practical relevance, for which CA is not needed. On the other hand, it is usually much easier to first obtain asymptotic statements which requires CA, and then improve upon them. Furthermore we will show that CA can guide us in the right direction to find good finitary prior probabilities.

Another principle which has received much less attention than CA but is equally if not more important is that of Cournot [Cou43, Sha06]: An event of probability (close to) zero singled out in advance is physically impossible; or conversely, an event of probability 1 will physically happen for sure. In short: zero probability means impossibility. The history of the semantics of probability is stony [Fin73]. Cournot’s “forgotten” principle is one way of giving meaning to probabilistic statements like, “the relative frequency of heads of a fair coin converges to  $1/2$  with probability 1”. The contraposition of Cournot is that one must assign non-zero probability to possible events. If “events” are described by sentences and “possible” means it is possible to satisfy these sentences, i.e. they possess a model, then we arrive at the strong Cournot principle that satisfiable sentences should be assigned non-zero probability (Definitions 25 and 35). This condition has been appropriately called ‘non-dogmatic’ in [GS82]. As long as something is not proven false, there is a (small) chance it is valid in the intended interpretation. This non-dogmatism is crucial in Bayesian inductive reasoning, since no evidence (however strong) can increase a zero prior belief to a non-zero posterior belief [RH11]. The Gaifman condition is inconsistent with the strong Cournot principle (Example 43), but consistent with a weaker version (Definition 26). Probabilities that are Gaifman and (plain, not strong) Cournot allow learning in the limit (Theorem 27 and Corollary 64).

A standard way to construct (general / Cournot / Gaifman) probabilities on sentences is to construct (general / non-dogmatic / separating) probabilities on interpretations, and then transfer them to sentences (Propositions 29, 32, and 38). At the same time we give model-theoretic characterizations of the Gaifman condition (Corollary 34) and the Cournot condition (Definition 37). In Section 5, we give a particularly simple construction of a probability that is Cournot and Gaifman (Theorem 40) and a complete characterization of general/Cournot/Gaifman probabilities (Theorems 50 and 52 and Corollary 53). We also give various examples of (strong)

(non)Cournot and/or Gaifman probabilities and (non)separating interpretations for countable domains (Examples 46, 47, and 48) and finite domains (Examples 42, 43, 44, 45).

Section 7 considers the important practical situation of whether a real-valued function on a set of sentences can be extended to a probability on all sentences; a necessary and sufficient condition is given for this, as is a method for determining such probabilities using minimum relative entropy introduced in Section 6. Prior knowledge and data constrain our (belief) probabilities in various ways, which we need to take into account when constructing probabilities. Prior knowledge is usually given in the form of probabilities on sentences like “the coin has head probability  $1/2$ ”, or facts like “all electrons have the same charge”, or non-logical axioms like “there are infinitely many natural numbers”. They correspond to requiring their probability to be  $1/2$ , extremely close to 1, and 1, respectively. It is therefore necessary to be able to go from probabilities on sentences to probability on interpretations (Proposition 31). This allows us to prove various necessary and sufficient conditions under which such partial probability specifications can be completed and what properties they have (Propositions 57 and 60). In particular we show that hierarchical probabilistic knowledge (Definitions 61) is always probabilistically consistent (Proposition 63). Further, seldom does knowledge constrain the probability on all sentences to be uniquely determined. In this case it is natural to choose a probability that is least dogmatic or biased [Nil86, Wil08a]. The minimum relative entropy (Definition 54) principle can be used to construct such a unique minimally more informative probability that is consistent with our prior knowledge (Definition 55 and Propositions 56 and 57).

Section 8 is a brief outlook on how the developed theory might be used and approximated in autonomous reasoning agents. In particular, certain knowledge, learning in the limit (64), the infamous black raven paradox, and the Monty Hall problem are discussed, but only briefly. The paper ends with a more detailed discussion in Section 9 of the broader context and motivation of this work, as well as related results in the literature, the outline of a framework for probabilistic reasoning and modeling in higher-order logic, and future research directions.

While some of the results presented in this paper are known in the first-order case and their extension to the higher-order case is straightforward, it nevertheless seems useful to provide a survey of this material (with proofs included). Also, many beautiful ideas in the long and technical paper by Gaifman [GS82] deserve wider attention than they have received. We hope our exposition helps to rectify this situation.

## 2 Logic

We review here a standard formulation of higher-order logic [And02] that is based on Church’s simple theory of types [Chu40]. Other references on higher-order logic include [Llo03, Far08, vBD83, Lei94, Sha01]. Some discussion of the interesting history of the simple theory of types is given in [And02, Far08].

The best way to think about higher-order logic is that it is the formalization of everyday informal mathematics: whatever mathematical description one might give of some situation, the formalization of that situation in higher-order logic is likely to be a straightforward translation of the informal description. In particular, higher-order logic provides a suitable foundation for mathematics itself which has several advantages over more traditional approaches that are based on axiomatizing sets in first-order logic. Furthermore, higher-order logic is the logical formalism of choice for much of theoretical computer science and also applications areas such as software and hardware verification. For a convincing account of the advantages of higher-order

over first-order logic in computer science, see [Far08].

The logic presented here differs in a minor way from that in [And02] in that we omit the description operator  $\iota$ , for reasons that are discussed later. All the results from [And02] that are used here also hold for the logic with  $\iota$  omitted, by obvious changes to their proofs. In addition the notation for the logic used here differs somewhat from that in [And02], but the correspondences will always be clear. There are also a few differences in terminology here compared to [And02] that are noted along the way.

We begin with the definition of a type.

**Definition 1 (type  $\alpha$ )** *A type is defined inductively as follows.*

1.  $o$  is a type.
2.  $\iota$  is a type.
3. If  $\alpha$  and  $\beta$  are types, then  $\alpha \rightarrow \beta$  is a type.

In this definition,  $o$  is the type of the truth values,  $\iota$  is the type of individuals, and  $\alpha \rightarrow \beta$  is the type of functions from elements of type  $\alpha$  to elements of type  $\beta$ . We use the convention that  $\rightarrow$  is right associative. So, for example, when we write  $\alpha \rightarrow \beta \rightarrow \gamma \rightarrow \kappa$  we mean  $\alpha \rightarrow (\beta \rightarrow (\gamma \rightarrow \kappa))$ . A *function type* is a type of the form  $\alpha \rightarrow \beta$ , for some  $\alpha$  and  $\beta$ .

There is a denumerable list of variables of each type. The logical constants are  $=_{\alpha \rightarrow \alpha \rightarrow o}$ , for each type  $\alpha$ . The denotation of equality  $=_{\alpha \rightarrow \alpha \rightarrow o}$  is the identity relation between individuals of type  $\alpha$ . In addition, there may be other non-logical constants of various types. The *alphabet* is the set of all variables and constants.

Next comes the definition of a term.

**Definition 2 (term  $t$ )** *A term, together with its type, is defined inductively as follows.*

1. A variable of type  $\alpha$  is a term of type  $\alpha$ .
2. A constant of type  $\alpha$  is a term of type  $\alpha$ .
3. If  $t_\beta$  is a term of type  $\beta$  and  $x_\alpha$  a variable of type  $\alpha$ , then  $\lambda x_\alpha. t_\beta$  is a term of type  $\alpha \rightarrow \beta$ .
4. If  $s_{\alpha \rightarrow \beta}$  is a term of type  $\alpha \rightarrow \beta$  and  $t_\alpha$  a term of type  $\alpha$ , then  $(s_{\alpha \rightarrow \beta} t_\alpha)$  is a term of type  $\beta$ .

A formula is a term of type  $o$ . A closed term is a term with no free variables. A sentence is a closed formula. A theory is a set of formulas.

If the set of non-logical constants is countable, then the set of terms is denumerable. As shown in [And02, p.212], using equality, it is easy to define  $\top_o$  (truth),  $\perp_o$  (falsity),  $\wedge_{o \rightarrow o \rightarrow o}$  (conjunction),  $\vee_{o \rightarrow o \rightarrow o}$  (disjunction),  $\neg_{o \rightarrow o}$  (negation),  $\forall x_\alpha. t_o$  (universal quantification), and  $\exists x_\alpha. t_o$  (existential quantification). The axioms for the logic are as follows [And02, p.213]:

**Axiom 3 (logical axioms)**

1. *Truth values:*  $(g_{o \rightarrow o} \top_o) \wedge (g_{o \rightarrow o} \perp_o) = \forall x_o. (g_{o \rightarrow o} x_o)$
2. *Leibniz' law:*  $(x_\alpha = y_\alpha) \rightarrow ((h_{\alpha \rightarrow o} x_\alpha) = (h_{\alpha \rightarrow o} y_\alpha))$
3. *Extensionality:*  $(f_{\alpha \rightarrow \beta} = g_{\alpha \rightarrow \beta}) = \forall x_\alpha. ((f_{\alpha \rightarrow \beta} x_\alpha) = (g_{\alpha \rightarrow \beta} x_\alpha))$
4.  *$\beta$ -reduction:*  $(\lambda \mathbf{x}_\alpha. \mathbf{t}_\beta \mathbf{s}_\alpha) = \mathbf{t}_\beta \{\mathbf{x}_\alpha / \mathbf{s}_\alpha\}$  (provided that  $\mathbf{s}_\alpha$  is free for  $\mathbf{x}_\alpha$  in  $\mathbf{t}_\beta$ )

In the above,  $g_{o \rightarrow o}, \dots$  are variables of the indicated type,  $\mathbf{x}_\alpha$  is a syntactical variable for variables of type  $\alpha$ , and  $\mathbf{t}_\beta, \dots$  are syntactical variables for terms of the indicated type. Also  $\mathbf{t}_\beta\{\mathbf{x}_\alpha/\mathbf{s}_\alpha\}$  is the result of simultaneously substituting  $\mathbf{s}_\alpha$  for all free occurrences of  $\mathbf{x}_\alpha$  in  $\mathbf{t}_\beta$ .

Axiom (1) expresses the idea the truth and falsity are the only truth values; Axioms (2) (for each type  $\alpha$ ) express a basic property of equality; Axioms (3) (for each type  $\alpha \rightarrow \beta$ ) are the axioms of extensionality; and Axiom schemata (4) is the axiom for  $\beta$ -reduction.

Here is the single rule of inference [And02, p.213]:

**Rule 4 (rule of inference; equality substitution)** *From  $\mathbf{t}_o$  and  $\mathbf{s}_\alpha = \mathbf{r}_\alpha$ , infer the result of replacing one occurrence of  $\mathbf{s}_\alpha$  in  $\mathbf{t}_o$  by an occurrence of  $\mathbf{r}_\alpha$ , provided that the occurrence of  $\mathbf{s}_\alpha$  in  $\mathbf{t}_o$  is not (an occurrence of a variable) immediately preceded by a  $\lambda$ .*

The logic also has an equational reasoning system that has been used as the computational basis for a functional logic programming language [Llo03, NL09, NLU08, LN11].

In the following, to simplify the notation, we usually omit the type subscripts on terms; the type of a term will always either be unimportant or clear from the context. We use  $\varphi, \chi, \psi$  for sentences and sometimes for formulas, and  $t, r, s$  for terms. With this notation,  $\forall x.\varphi \equiv [\lambda x.\varphi = \lambda x.\top]$  and  $\exists x.\varphi \equiv [\lambda x.\varphi \neq \lambda x.\perp]$ .

The logic includes Church's  $\lambda$ -calculus: a term of the form  $\lambda x.t$  is an abstraction and a term of the form  $(s\ t)$  is an application.

The logic is given a conventional Henkin semantics [Hen50].

**Definition 5 (frame  $\{\mathcal{D}_\alpha\}_\alpha$ )** *A frame is a collection  $\{\mathcal{D}_\alpha\}_\alpha$  of non-empty sets, one for each type  $\alpha$ , satisfying the following conditions.*

1.  $\mathcal{D}_o = \{\top, \text{F}\}$ .
2.  $\mathcal{D}_{\beta \rightarrow \gamma}$  is some collection of functions from  $\mathcal{D}_\beta$  to  $\mathcal{D}_\gamma$ .

For each type  $\alpha$ ,  $\mathcal{D}_\alpha$  is called a domain.

The members of  $\mathcal{D}_o$  are called the *truth values* and the members of  $\mathcal{D}_i$  are called *individuals*.

**Definition 6 (valuation  $V$ )** *Given a frame  $\{\mathcal{D}_\alpha\}_\alpha$ , a valuation  $V$  is a function that maps each constant having type  $\alpha$  to an element of  $\mathcal{D}_\alpha$  such that  $V(=_{\alpha \rightarrow \alpha \rightarrow o})$  is the function from  $\mathcal{D}_\alpha$  into  $\mathcal{D}_{\alpha \rightarrow o}$  defined by*

$$V(=_{\alpha \rightarrow \alpha \rightarrow o})\ x\ y = \begin{cases} \top & \text{if } x = y \\ \text{F} & \text{otherwise,} \end{cases}$$

for  $x, y \in \mathcal{D}_\alpha$ .

**Definition 7 (variable assignment  $\nu$ )** *A variable assignment  $\nu$  with respect to a frame  $\{\mathcal{D}_\alpha\}_\alpha$  is a function that maps each variable of type  $\alpha$  to an element of  $\mathcal{D}_\alpha$ .*

An interpretation can now be defined.

**Definition 8 (interpretation  $\langle \{\mathcal{D}_\alpha\}_\alpha, V \rangle$ )** *A pair  $I \equiv \langle \{\mathcal{D}_\alpha\}_\alpha, V \rangle$  is an interpretation if there is a function  $\mathcal{V}$  such that, for each variable assignment  $\nu$  and for each term  $t$  of type  $\alpha$ ,  $\mathcal{V}(t, I, \nu) \in \mathcal{D}_\alpha$  and the following conditions are satisfied.*

1.  $\mathcal{V}(x, I, \nu) = \nu(x)$ , where  $x$  is a variable.

2.  $\mathcal{V}(C, I, \nu) = V(C)$ , where  $C$  is a constant.
3.  $\mathcal{V}(\lambda x.s, I, \nu) =$  the function whose value for each  $d \in \mathcal{D}_\beta$  is  $\mathcal{V}(s, I, \nu')$ , where  $\lambda x.s$  has type  $\beta \rightarrow \gamma$  and  $\nu'$  is  $\nu$  except  $\nu'(x) = d$ .
4.  $\mathcal{V}((r\ s), I, \nu) = \mathcal{V}(r, I, \nu)(\mathcal{V}(s, I, \nu))$ .

If  $\langle \{\mathcal{D}_\alpha\}_\alpha, V \rangle$  is an interpretation, then the function  $\mathcal{V}$  is uniquely defined.  $\mathcal{V}(t, I, \nu)$  is called the *denotation* of  $t$  with respect to  $I$  and  $\nu$ . If  $t$  is a closed term, then  $\mathcal{V}(t, I, \nu)$  is independent of  $\nu$  and we write it as  $\mathcal{V}(t, I)$ . Not every pair  $\langle \{\mathcal{D}_\alpha\}_\alpha, V \rangle$  is an interpretation; to be an interpretation, every term must have a denotation with respect to each variable assignment.

What is called an interpretation here is called a *general model* in [And02], following Henkin. In [And02], a general model is called a *standard model* if, for each  $\alpha$  and  $\beta$ ,  $\mathcal{D}_{\alpha \rightarrow \beta}$  is the set of all functions from  $\mathcal{D}_\alpha$  to  $\mathcal{D}_\beta$ . Moving from standard models to general models was the crucial step that allowed Henkin to prove the completeness of the logic [Hen50].

**Definition 9 (satisfiable)** Let  $t$  be a formula,  $I \equiv \langle \{\mathcal{D}_\alpha\}_\alpha, V \rangle$  an interpretation, and  $\nu$  a variable assignment with respect to  $\{\mathcal{D}_\alpha\}_\alpha$ .

1.  $\nu$  satisfies  $t$  in  $I$  if  $\mathcal{V}(t, I, \nu) = \top$ .
2.  $t$  is satisfiable in  $I$  if there is a variable assignment which satisfies  $t$  in  $I$ .
3.  $t$  is valid in  $I$  if every variable assignment satisfies  $t$  in  $I$ .
4.  $t$  is valid if  $t$  is valid in every interpretation.
5. A model for a theory is an interpretation in which each formula in the theory is valid.

**Definition 10 (consistency)** A theory is consistent if  $\perp$  cannot be derived from the theory.

**Definition 11 (logical consequence)** A formula  $t$  is a logical consequence of a theory if  $t$  is valid in every model of the theory.

We will have need for a particular class of interpretations, defined as follows.

**Definition 12 (separating interpretation/model)** An interpretation  $I$  for an alphabet is separating if, for every pair  $r, s$  of closed terms of the same function type, say,  $\alpha \rightarrow \beta$ , such that  $\mathcal{V}(r, I) \neq \mathcal{V}(s, I)$ , there exists a closed term  $t$  of type  $\alpha$  such that  $\mathcal{V}((r\ t), I) \neq \mathcal{V}((s\ t), I)$ .

A separating model is a separating interpretation that is a model (for some set of formulas).

We emphasize that, in the definition of a separating interpretation, the closed term  $t$  is formed only from symbols *in the given alphabet*. Intuitively, an interpretation is separating if, for every pair  $r, s$  of closed terms of the same type  $\alpha \rightarrow \beta$ , whose respective denotations in the interpretation are different, there exists a closed term  $t$  of type  $\alpha$  for which the respective denotations in the interpretation of  $(r\ t)$  and  $(s\ t)$  are different. Thus, in a separating interpretation, closed terms that have distinct functions as denotations must be distinct on an argument in the domain that is the denotation of some closed term using the given alphabet and thus is ‘accessible’ or ‘nameable’ via that term.

The concept of a separating interpretation is closely related to the concept of an extensionally complete theory that plays a crucial part in the proof of completeness [And02, p.248].

**Definition 13 (extensionally complete)** A set  $S$  of sentences is extensionally complete if, for every pair  $r, s$  of closed terms of the same function type, say,  $\alpha \rightarrow \beta$ , there exists a closed term  $t$  of type  $\alpha$  such that  $r \neq s \rightarrow (r\ t) \neq (s\ t)$  is derivable from  $S$ .

A connection with separating interpretations is provided by the following result.

**Proposition 14 (extensionally complete  $\Rightarrow$  separating)** *Every model of an extensionally complete set of sentences is separating.*

**Proof.** Let  $S$  be a set of sentences that is extensionally complete and  $I$  be a model for  $S$ . Suppose that  $r, s$  is a pair of closed terms of the same function type, say,  $\alpha \rightarrow \beta$ , such that  $\mathcal{V}(r, I) \neq \mathcal{V}(s, I)$ . By extensional completeness, there exists a closed term  $t$  such that  $r \neq s \rightarrow (r\ t) \neq (s\ t)$  is derivable from  $S$ . Since  $I$  is a model for  $S$  and the proof system is sound, it follows that  $\mathcal{V}((r\ t), I) \neq \mathcal{V}((s\ t), I)$ . Hence  $I$  is separating. ■

Now we show that, if we are willing to expand the alphabet, any set of sentences having a model also has a separating model in an expanded alphabet.

**Proposition 15 (existence of separating models)** *If a set  $S$  of sentences has a model, then there exists an alphabet that includes the original alphabet and an interpretation based on the expanded alphabet which is a separating model for  $S$ .*

**Proof.** Since  $S$  has a model,  $S$  is consistent. By [And02, Theorem 5500], there is an expansion of the original alphabet and a set  $T$  of sentences such that  $S \subseteq T$ ,  $T$  is consistent, and  $T$  is extensionally complete in the expanded alphabet. Since  $T$  is consistent, by Henkin's Theorem [And02, Theorem 5501], it has a model (based on the expanded alphabet). By Proposition 14, this model must be a separating one, and it is also a model for  $S$ . ■

The most important property of the logic that we will need is compactness [And02, Theorem 5503].

**Theorem 16 (compactness)** *If every finite subset of a set  $S$  of sentences has a model, then  $S$  has a model.*

In fact, most of the development in the paper can be carried out in any logic that has the compactness property.

While the version of higher-order logic introduced in this section generally provides much more direct and succinct formalisations than first-order logic, for practical applications a number of extensions are highly desirable. Some of these extensions are nothing more than abbreviations, such as those used to introduce the connectives and quantifiers, and some are deeper. These extensions include many-sortedness, which allows more than one domain of individuals; tuples and product types; and type constructors and polymorphism. The logic of [Llo03], which is also used in [NL09, NLU08], includes all these extensions. These and other extensions are discussed in [Far08].

### 3 Probabilities on Sentences

We now define probabilities on sentences. They are not probabilities in the conventional sense of probability theory (on  $\sigma$ -algebras); however, a connection between probabilities on sentences and (conventional) probabilities on a  $\sigma$ -algebra on the set of interpretations will be made below.



**Definition 17 (probability on sentences)** Let  $\mathcal{S}$  be the set of all sentences (for some alphabet). A probability (on sentences) is a non-negative function  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  satisfying the following conditions:

1. If  $\varphi$  is valid, then  $\mu(\varphi) = 1$ .
2. If  $\neg(\varphi \wedge \psi)$  is valid, then  $\mu(\varphi \vee \psi) = \mu(\varphi) + \mu(\psi)$ .

For a sentence  $\psi$ , where  $\mu(\psi) > 0$ , one can define the conditional probability  $\mu(\cdot|\psi)$  by

$$\mu(\varphi|\psi) = \frac{\mu(\varphi \wedge \psi)}{\mu(\psi)},$$

for each sentence  $\varphi$ .

A probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  on sets of sentences has the following intended meaning:

For a sentence  $\varphi$ ,  $\mu(\varphi)$  is the degree of belief that  $\varphi$  is true.

**Definition 18 (pairwise disjoint sentences)** The sentences  $\varphi_1, \dots, \varphi_n$  are pairwise disjoint if, for each  $i, j = 1, \dots, n$  such that  $i \neq j$ ,  $\neg(\varphi_i \wedge \varphi_j)$  is valid.

**Proposition 19 (properties of probability on sentences)** Let  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  be a probability on sentences. Then the following hold:

1.  $\mu(\neg\varphi) = 1 - \mu(\varphi)$ , for each  $\varphi \in \mathcal{S}$ .
2.  $\mu(\varphi) \leq 1$ , for each  $\varphi \in \mathcal{S}$ .
3. If  $\varphi$  is unsatisfiable, then  $\mu(\varphi) = 0$ .
4. If  $\varphi \rightarrow \psi$  is valid, then  $\mu(\varphi) \leq \mu(\psi)$ .
5. If  $\varphi = \psi$  is valid, then  $\mu(\varphi) = \mu(\psi)$ .
6. If  $\{\varphi_i\}_{i=1}^n$  is a finite subset of pairwise disjoint sentences in  $\mathcal{S}$ , then  $\mu(\bigvee_{i=1}^n \varphi_i) = \sum_{i=1}^n \mu(\varphi_i)$ .
7. If  $\{\varphi_i\}_{i=1}^n$  is a finite subset of  $\mathcal{S}$ , then  $\mu(\bigvee_{i=1}^n \varphi_i) \leq \sum_{i=1}^n \mu(\varphi_i)$ .
8. The following are equivalent:
  - (a) For each  $\varphi \in \mathcal{S}$ ,  $\mu(\varphi) = 1$  implies  $\varphi$  is valid.
  - (b) For each  $\varphi \in \mathcal{S}$ ,  $\mu(\varphi) = 0$  implies  $\varphi$  is unsatisfiable.
9. If  $\mu(\psi) > 0$ , then  $\mu(\cdot|\psi)$  is a probability.
10.  $\mu(\varphi \vee \psi) + \mu(\varphi \wedge \psi) = \mu(\varphi) + \mu(\psi)$ .

**Proof.** The proof is elementary and standard, and only included for completeness.

1. Since  $\neg(\varphi \wedge \neg\varphi)$  is valid,  $\mu(\varphi \vee \neg\varphi) = \mu(\varphi) + \mu(\neg\varphi)$ . Also, since  $\varphi \vee \neg\varphi$  is valid,  $\mu(\varphi \vee \neg\varphi) = 1$ . Thus  $\mu(\neg\varphi) = 1 - \mu(\varphi)$ .
2. Since  $1 - \mu(\varphi) = \mu(\neg\varphi) \geq 0$ , we have that  $\mu(\varphi) \leq 1$ .
3. Note that  $\varphi$  is unsatisfiable iff  $\neg\varphi$  is valid. Thus  $\mu(\neg\varphi) = 1 - \mu(\varphi) = 1$ , so that  $\mu(\varphi) = 0$ .
4. Note first that  $\varphi \rightarrow \psi$  is valid iff  $\neg(\varphi \wedge \neg\psi)$  is valid. Thus  $\mu(\varphi \vee \neg\psi) = \mu(\varphi) + \mu(\neg\psi) = \mu(\varphi) + 1 - \mu(\psi)$ . Hence  $\mu(\varphi) = \mu(\psi) + \mu(\varphi \vee \neg\psi) - 1 \leq \mu(\psi)$ .
5. This follows immediately from Part 4.

6. The proof is by induction on  $n$ . When  $n = 1$  the result is obvious. Assume now the result is true for  $n - 1$ . Note that  $\bigwedge_{i=2}^n \neg(\varphi_1 \wedge \varphi_i)$  is valid and so  $\neg(\varphi_1 \wedge \bigvee_{i=2}^n \varphi_i)$  is valid. Then

$$\begin{aligned}
& \mu(\bigvee_{i=1}^n \varphi_i) \\
&= \mu(\varphi_1 \vee \bigvee_{i=2}^n \varphi_i) \\
&= \mu(\varphi_1) + \mu(\bigvee_{i=2}^n \varphi_i) && [\neg(\varphi_1 \wedge \bigvee_{i=2}^n \varphi_i) \text{ is valid}] \\
&= \mu(\varphi_1) + \sum_{i=2}^n \mu(\varphi_i) && [\text{induction hypothesis}] \\
&= \sum_{i=1}^n \mu(\varphi_i).
\end{aligned}$$

7. The proof is by induction on  $n$ . When  $n = 1$  the result is obvious. Assume now the result is true for  $n - 1$ . Then

$$\begin{aligned}
& \mu(\bigvee_{i=1}^n \varphi_i) \\
&= \mu((\varphi_1 \wedge \neg \bigvee_{i=2}^n \varphi_i) \vee \bigvee_{i=2}^n \varphi_i) \\
&= \mu(\varphi_1 \wedge \neg \bigvee_{i=2}^n \varphi_i) + \mu(\bigvee_{i=2}^n \varphi_i) \\
&\leq \mu(\varphi_1) + \sum_{i=2}^n \mu(\varphi_i) && [\text{Part 4 and induction hypothesis}] \\
&= \sum_{i=1}^n \mu(\varphi_i).
\end{aligned}$$

8. Suppose that, for each  $\varphi \in \mathcal{S}$ ,  $\mu(\varphi) = 1$  implies  $\varphi$  is valid. Now let  $\psi \in \mathcal{S}$  satisfy  $\mu(\psi) = 0$ . By Part 1,  $\mu(\neg\psi) = 1$ . Thus  $\neg\psi$  is valid and so  $\psi$  is unsatisfiable.

Conversely, suppose that, for each  $\varphi \in \mathcal{S}$ ,  $\mu(\varphi) = 0$  implies  $\varphi$  is unsatisfiable. Now let  $\psi \in \mathcal{S}$  satisfy  $\mu(\psi) = 1$ . By Part 1,  $\mu(\neg\psi) = 0$ . Thus  $\neg\psi$  is unsatisfiable and so  $\psi$  is valid.

9. Suppose that  $\varphi$  is valid. Then  $\mu(\varphi|\psi) = \frac{\mu(\varphi \wedge \psi)}{\mu(\psi)} = \frac{\mu(\psi)}{\mu(\psi)} = 1$ .

Suppose that  $\neg(\varphi \wedge \chi)$  is valid. Then

$$\begin{aligned}
& \mu(\varphi \vee \chi|\psi) \\
&= \mu((\varphi \vee \chi) \wedge \psi) / \mu(\psi) \\
&= \mu((\varphi \wedge \psi) \vee (\chi \wedge \psi)) / \mu(\psi) \\
&= [\mu(\varphi \wedge \psi) + \mu(\chi \wedge \psi)] / \mu(\psi) && [\neg((\varphi \wedge \psi) \wedge (\chi \wedge \psi)) \text{ is valid}] \\
&= \mu(\varphi|\psi) + \mu(\chi|\psi).
\end{aligned}$$

Thus  $\mu(\cdot|\psi)$  is a probability.

10. Let  $\chi := \neg\varphi \wedge \psi$ . Then

$$\begin{aligned}
& \mu(\varphi \vee \psi) + \mu(\varphi \wedge \psi) \\
&= \mu(\varphi \vee \chi) + \mu(\varphi \wedge \psi) && [\text{elementary logic}] \\
&= \mu(\varphi) + \mu(\chi) + \mu(\varphi \wedge \psi) && [\neg(\varphi \wedge \chi) \text{ is valid and Def. 17.2}] \\
&= \mu(\varphi) + \mu(\chi \vee (\varphi \wedge \psi)) && [\neg(\chi \wedge (\varphi \wedge \psi)) \text{ is valid and Def. 17.2}] \\
&= \mu(\varphi) + \mu(\psi) && [\text{elementary logic}]
\end{aligned}$$

■

Next we introduce Gaifman probabilities.

**Definition 20 (Gaifman probability)** Let  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  be a probability on sentences. Then  $\mu$  is Gaifman if

$$\mu(r = s) = \inf_{\{t_1, \dots, t_n\}} \mu\left(\bigwedge_{i=1}^n ((r \ t_i) = (s \ t_i))\right),$$

for every pair  $r$  and  $s$  of closed terms having the same function type, say,  $\alpha \rightarrow \beta$ , and where  $\{t_1, \dots, t_n\}$  ranges over all finite sets of closed terms of type  $\alpha$ .

**Proposition 21 (Gaifman probability)** Let  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  be a probability on sentences. Then the following are equivalent.

1.  $\mu$  is Gaifman.

$$2. \mu(r \neq s) = \sup_{\{t_1, \dots, t_n\}} \mu\left(\bigvee_{i=1}^n ((r \ t_i) \neq (s \ t_i))\right),$$

for every pair  $r$  and  $s$  of closed terms having the same function type, say,  $\alpha \rightarrow \beta$ , and where  $\{t_1, \dots, t_n\}$  ranges over all finite sets of closed terms of type  $\alpha$ .

$$3. \mu(\exists x. \varphi) = \sup_{\{t_1, \dots, t_n\}} \mu\left(\bigvee_{i=1}^n \varphi\{x/t_i\}\right),$$

for every formula  $\varphi$  having a single free variable  $x$  of type  $\alpha$ , say, and where  $\{t_1, \dots, t_n\}$  ranges over all finite sets of closed terms of type  $\alpha$ .

$$4. \mu(\forall x. \varphi) = \inf_{\{t_1, \dots, t_n\}} \mu\left(\bigwedge_{i=1}^n \varphi\{x/t_i\}\right),$$

for every formula  $\varphi$  having a single free variable  $x$  of type  $\alpha$ , say, and where  $\{t_1, \dots, t_n\}$  ranges over all finite sets of closed terms of type  $\alpha$ .

**Proof.** 1. implies 2. Suppose that the probability  $\mu$  is Gaifman. Then

$$\begin{aligned} & \mu(r \neq s) \\ &= 1 - \mu(r = s) \\ &= 1 - \inf_{\{t_1, \dots, t_n\}} \mu\left(\bigwedge_{i=1}^n ((r \ t_i) = (s \ t_i))\right) \\ &= 1 - \inf_{\{t_1, \dots, t_n\}} \mu(\neg \bigvee_{i=1}^n ((r \ t_i) \neq (s \ t_i))) \\ &= 1 - \inf_{\{t_1, \dots, t_n\}} (1 - \mu(\bigvee_{i=1}^n ((r \ t_i) \neq (s \ t_i)))) \\ &= \sup_{\{t_1, \dots, t_n\}} \mu(\bigvee_{i=1}^n ((r \ t_i) \neq (s \ t_i))). \end{aligned}$$

Hence 2. holds.

2. implies 3. Suppose that 2. holds. Then

$$\begin{aligned} & \mu(\exists x. \varphi) \\ &= \mu(\lambda x. \varphi \neq \lambda x. F) \\ &= \sup_{\{t_1, \dots, t_n\}} \mu(\bigvee_{i=1}^n ((\lambda x. \varphi \ t_i) \neq (\lambda x. F \ t_i))) \\ &= \sup_{\{t_1, \dots, t_n\}} \mu(\bigvee_{i=1}^n \varphi\{x/t_i\}). \end{aligned}$$

Hence 3. holds.

3. implies 4. Suppose that 3. holds. Then

$$\begin{aligned}
& \mu(\forall x. \varphi) \\
&= \mu(\neg \exists x. \neg \varphi) \\
&= 1 - \mu(\exists x. \neg \varphi) \\
&= 1 - \sup_{\{t_1, \dots, t_n\}} \mu(\bigvee_{i=1}^n \neg \varphi\{x/t_i\}) \\
&= 1 - \sup_{\{t_1, \dots, t_n\}} \mu(\neg \bigwedge_{i=1}^n \varphi\{x/t_i\}) \\
&= 1 - \sup_{\{t_1, \dots, t_n\}} (1 - \mu(\bigwedge_{i=1}^n \varphi\{x/t_i\})) \\
&= \inf_{\{t_1, \dots, t_n\}} \mu(\bigwedge_{i=1}^n \varphi\{x/t_i\}).
\end{aligned}$$

Hence 4. holds.

4. implies 1. Suppose that 4. holds. Then

$$\begin{aligned}
& \mu(r = s) \\
&= \mu(\forall x. ((r\ x) = (s\ x))) && [\text{Axioms of Extensionality}] \\
&= \inf_{\{t_1, \dots, t_n\}} \mu(\bigwedge_{i=1}^n ((r\ t_i) = (s\ t_i))) \\
&= \inf_{\{t_1, \dots, t_n\}} \mu(\bigwedge_{i=1}^n ((r\ t_i) = (s\ t_i))).
\end{aligned}$$

Hence 1. holds. ■

**Proposition 22 (limits for countable alphabet)** *Let the alphabet be countable,  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  a probability on sentences, and  $\varphi$  a formula having a single free variable  $x$  of type  $\alpha$ .*

$$\begin{aligned}
1. \quad & \sup_{\{t_1, \dots, t_n\}} \mu(\bigvee_{i=1}^n \varphi\{x/t_i\}) = \lim_{n \rightarrow \infty} \mu(\bigvee_{i=1}^n \varphi\{x/t_i\}) \\
2. \quad & \inf_{\{t_1, \dots, t_n\}} \mu(\bigwedge_{i=1}^n \varphi\{x/t_i\}) = \lim_{n \rightarrow \infty} \mu(\bigwedge_{i=1}^n \varphi\{x/t_i\}),
\end{aligned}$$

where, on the LHS,  $\{t_1, \dots, t_n\}$  ranges over all finite sets of closed terms of type  $\alpha$  and, on the RHS,  $t_1, t_2, \dots$  is an enumeration of all closed terms of type  $\alpha$ .

**Proof.** Since the alphabet is countable, the set of all closed terms of type  $\alpha$  is countable and hence can be enumerated.

1. Let  $\{t'_1, \dots, t'_m\}$  be a subset of closed terms of type  $\alpha$ . Let  $n$  be sufficiently large so that each  $t'_j$ , for  $j = 1, \dots, m$ , appears in the enumeration  $t_1, \dots, t_n$  of the first  $n$  terms of an enumeration of all closed terms of type  $\alpha$ .

Then  $\bigvee_{j=1}^m \varphi\{x/t'_j\} \rightarrow \bigvee_{i=1}^n \varphi\{x/t_i\}$  is valid, so that

$$\mu(\bigvee_{j=1}^m \varphi\{x/t'_j\}) \leq \mu(\bigvee_{i=1}^n \varphi\{x/t_i\}),$$

by Proposition 19.4. By first taking the supremum on the RHS and then the supremum on the LHS we get

$$\sup_{\{t'_1, \dots, t'_m\}} \mu(\bigvee_{j=1}^m \varphi\{x/t'_j\}) \leq \sup_n \mu(\bigvee_{i=1}^n \varphi\{x/t_i\}).$$

Conversely we have

$$\sup_{\{t'_1, \dots, t'_m\}} \mu(\bigvee_{j=1}^m \varphi\{x/t'_j\}) \geq \mu(\bigvee_{i=1}^n \varphi\{x/t_i\}).$$

since the sup on the LHS includes  $\{t_1, \dots, t_n\}$ . Now taking the limit  $n \rightarrow \infty$  and combining both inequalities gives equality. Proposition 19.4 gives that  $\mu(\varphi \vee \psi) \geq \mu(\varphi)$ ; hence  $\mu(\bigvee_{i=1}^n \varphi\{x/t_i\})$  is monotone non-decreasing in  $n$ , which allows the replacement of  $\sup_n$  by  $\lim_{n \rightarrow \infty}$ .

2. The proof is similar. ■

We can reduce the class of terms that is necessary to “browse” through even further, by considering only one term from each equivalence class, where two terms  $t$  and  $t'$  are equivalent iff  $t = t'$  is valid.

**Proposition 23 (Gaifman for countable alphabet)** *Let the alphabet be countable and  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  a probability on sentences. Then the following are equivalent.*

1.  $\mu$  is Gaifman.

$$2. \mu(r = s) = \lim_{n \rightarrow \infty} \mu\left(\bigwedge_{i=1}^n ((r \ t_i) = (s \ t_i))\right),$$

for every pair  $r$  and  $s$  of closed terms having the same function type, say,  $\alpha \rightarrow \beta$ , and where  $t_1, t_2, \dots$  is an enumeration of all closed terms of type  $\alpha$ .

$$3. \mu(r \neq s) = \lim_{n \rightarrow \infty} \mu\left(\bigvee_{i=1}^n ((r \ t_i) \neq (s \ t_i))\right),$$

for every pair  $r$  and  $s$  of closed terms having the same function type, say,  $\alpha \rightarrow \beta$ , and where  $t_1, t_2, \dots$  is an enumeration of all closed terms of type  $\alpha$ .

$$4. \mu(\exists x. \varphi) = \lim_{n \rightarrow \infty} \mu\left(\bigvee_{i=1}^n \varphi\{x/t_i\}\right),$$

for every formula  $\varphi$  having a single free variable  $x$  of type  $\alpha$ , say, and where  $t_1, t_2, \dots$  is an enumeration of all closed terms of type  $\alpha$ .

$$5. \mu(\forall x. \varphi) = \lim_{n \rightarrow \infty} \mu\left(\bigwedge_{i=1}^n \varphi\{x/t_i\}\right),$$

for every formula  $\varphi$  having a single free variable  $x$  of type  $\alpha$ , say, and where  $t_1, t_2, \dots$  is an enumeration of all closed terms of type  $\alpha$ .

In each case, the enumeration  $t_1, t_2, \dots$  of closed terms of type  $\alpha$  can be reduced to one where a single representative is chosen from each equivalence class under the equivalence relation  $t$  and  $t'$  are equivalent if  $t = t'$  is valid.

**Proof.** Two terms  $t$  and  $t'$  are said to be equivalent iff  $t = t'$  is valid, which implies  $\varphi\{x/t\} = \varphi\{x/t'\}$  is valid. This allows us to relax in the proof of Proposition 22 ‘appears’ by ‘is equivalent to some term in’ and ‘includes’ by ‘includes a term equivalent to some term in’. Finally combine this with Proposition 21 and Definition 20. ■

While these forms of the Gaifman condition closely resemble the continuity condition (countable additivity (CA) axiom) in measure theory, we will see that CA over (general) interpretations is derived from the compactness theorem and not from the Gaifman condition (see Definition 28 and Proposition 30 in the next section). But the Gaifman condition confines probabilities to separating interpretations while preserving CA (Propositions 29 and 31).

**Example 24 (natural numbers *Nat*)** Consider the standard type *Nat* of natural numbers, as the type of individuals, and the usual Peano axioms. Let  $\underline{0}$  be the constant of type *Nat* whose denotation is the natural number 0, and  $\underline{n} \equiv S^n(\underline{0}) = (S (S (S \cdots (S \underline{0}))))$  be the term of type *Nat* whose denotation is the natural number  $n$ , where  $S$  is a constant of type  $\text{Nat} \rightarrow \text{Nat}$  whose denotation is the successor function. In practice one usually defines denumerably many constants  $\underline{1}, \underline{2}, \underline{3}, \dots$ , one for each natural number, directly. Further, let  $+, \times : \text{Nat} \rightarrow \text{Nat} \rightarrow \text{Nat}$  be functions with their usual axioms and meaning. Now there are many closed terms that represent the same natural number. For instance  $\underline{8}$ ,  $(\lambda x.x \underline{8})$ ,  $(\underline{3} + \underline{5})$ ,  $(\underline{2} \times \underline{4})$  are different terms, all having the number 8 as denotation. For type *Nat*, it is sufficient to choose  $t_n = \underline{n}$  in Proposition 23.4, and so the condition in Definition 20 (indeed) reduces to the one used by Gaifman [GS82].  $\diamond$

Of particular interest are probabilities that are strictly positive on satisfiable sentences since this is a desirable property of a prior. This suggests the following definition.

**Definition 25 (strongly Cournot probability)** A probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  is strongly Cournot if, for each  $\varphi \in \mathcal{S}$ ,  $\varphi$  is satisfiable implies  $\mu(\varphi) > 0$ .

By Part 8 of Proposition 19, a probability is strongly Cournot iff, for each  $\varphi \in \mathcal{S}$ ,  $\varphi$  is not valid implies  $\mu(\varphi) < 1$ , or, by contraposition,  $\mu(\varphi) = 1$  implies  $\varphi$  is valid. This is akin to Cournot's principle as discussed in the introduction that an event of probability 1 singled out in advance will happen for sure in the real world. We will see this general idea plays an important role for inductive inference.

However, the following weaker form of the Cournot principle will turn out to be more useful.

**Definition 26 (Cournot probability)** A probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  is Cournot if, for each  $\varphi \in \mathcal{S}$ ,  $\varphi$  has a separating model implies  $\mu(\varphi) > 0$ .

Clearly a strongly Cournot probability is Cournot. It will be the Cournot probabilities (not the strongly Cournot ones) that will be of most interest in the subsequent development. The major reasons for this are as follows. First, Theorem 40 below shows that, if the alphabet is countable, there exists a probability on sentences that is Cournot and Gaifman. Such a probability makes a good prior. Second, the Cournot and Gaifman conditions are necessary and sufficient to do learning in the limit of universal hypotheses as the following theorem shows and as discussed in more detail in Section 8.

**Theorem 27 (confirming universal hypotheses)** Let the alphabet be countable,  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  a probability on sentences,  $\varphi$  a formula having a single free variable  $x$  of some type  $\alpha$ ,  $t_1, t_2, \dots$  an enumeration of (representatives of) all closed terms of type  $\alpha$ . Then

$$\mu(\forall x.\varphi \mid \bigwedge_{i=1}^n \varphi\{x/t_i\}) \xrightarrow{n \rightarrow \infty} 1 \quad \Leftrightarrow \quad \mu(\bigwedge_{i=1}^n \varphi\{x/t_i\}) \xrightarrow{n \rightarrow \infty} \mu(\forall x.\varphi) > 0$$

If the left hand side (hence also the r.h.s.) holds, we say that  $\mu$  can confirm universal hypothesis  $\forall x.\varphi$ . It also holds that

$$\begin{array}{c} \mu \text{ can confirm all universal hypotheses} \\ \text{that have a separating model} \end{array} \Leftrightarrow \mu \text{ is Gaifman and Cournot}$$

$$\begin{aligned} \text{Proof. } (\text{top} \Leftarrow) & \lim_{n \rightarrow \infty} \mu(\forall x.\varphi \mid \bigwedge_{i=1}^n \varphi\{x/t_i\}) \\ &= \frac{\mu(\forall x.\varphi)}{\lim_{n \rightarrow \infty} \mu(\bigwedge_{i=1}^n \varphi\{x/t_i\})} & [\forall x.\varphi \rightarrow \bigwedge_{i=1}^n \varphi\{x/t_i\}] \\ &= \frac{\mu(\forall x.\varphi)}{\mu(\forall x.\varphi)} & [\bigwedge_{i=1}^n \varphi\{x/t_i\} \xrightarrow{n \rightarrow \infty} \mu(\forall x.\varphi)] \\ &= 1 & [\mu(\forall x.\varphi) > 0] \end{aligned}$$

(**top**  $\Rightarrow$ ) As can be seen from the  $\Leftarrow$  proof, if one or both of the conditions fail, then  $\mu(\forall x.\varphi \mid \bigwedge_{i=1}^n \varphi\{x/t_i\})$  does not converge to 1.

For the bottom  $\Leftrightarrow$  we abbreviate the statements

$$\begin{aligned} L(\varphi) &:= [\mu(\forall x.\varphi \mid \bigwedge_{i=1}^n \varphi\{x/t_i\}) \xrightarrow{n \rightarrow \infty} 1] \\ G(\varphi) &:= [\mu(\bigwedge_{i=1}^n \varphi\{x/t_i\}) \xrightarrow{n \rightarrow \infty} \mu(\forall x.\varphi)] \\ S(\varphi) &:= [\forall x.\varphi \text{ has a separating model}] \\ A(\varphi) &:= [\mu(\forall x.\varphi) > 0] \end{aligned}$$

In this notation, the  $\text{top} \Leftrightarrow$  reads  $L(\varphi)$  iff  $G(\varphi)$  and  $A(\varphi)$ .

(**bottom**  $\Leftarrow$ ) Assume  $\mu$  is Gaifman and Cournot and  $S(\varphi)$ . This implies  $G(\varphi)$  and  $A(\varphi)$ . By  $\text{top} \Leftarrow$  we get  $L(\varphi)$ . We have shown that for any  $\varphi$ , if  $\mu$  is Gaifman and Cournot, then  $S(\varphi)$  implies  $L(\varphi)$ .

(**bottom**  $\Rightarrow$ ) Case 1 [ $S(\varphi)$  is true] Then by assumption,  $L(\varphi)$ . Then by  $\text{top} \Rightarrow$  we get  $G(\varphi)$  and  $A(\varphi)$ . Note that every sentence  $\psi$  can be written as  $\psi = \forall x.\varphi$  with  $\varphi := [\psi \wedge (x = x)]$  being a formula having a single free variable  $x$ . Therefore,  $\mu(\psi) = \mu(\forall x.\varphi) > 0$  for all  $\psi$  that have a separating model. Hence  $\mu$  is Cournot.

Case 2 [ $S(\varphi)$  is false] That is,  $\forall x.\varphi$  has no separating model, therefore  $\neg\forall x.\varphi$  must have (at least one) separating model, say  $\hat{I}$ . Since  $\hat{I}$  is a *separating* model of  $\exists x.\neg\varphi$ , Definition 12 implies that there exists a closed term  $t$  such that  $\hat{I}$  is also a separating model of  $\chi := \neg\varphi\{x/t\}$ . Now

$$\begin{aligned} & \mu(\forall x.\varphi) + \mu(\chi) \\ &= \mu(\forall x.\varphi \vee \chi) & [\forall x.\varphi \text{ and } \chi \text{ are disjoint}] \\ &= \mu(\forall x.(\varphi \vee \chi)) & [x \text{ is not free in } \chi] \\ &= \lim_n \mu(\bigwedge_{i=1}^n (\varphi \vee \chi)\{x/t_i\}) & [\text{since } S(\varphi \vee \chi), \text{ Case 1 implies } G(\varphi \vee \chi)] \\ &= \lim_n \mu(\bigwedge_{i=1}^n \varphi\{x/t_i\} \vee \chi) & [x \text{ is not free in } \chi] \\ &= \lim_n \mu(\bigwedge_{i=1}^n \varphi\{x/t_i\}) + \mu(\chi) & [t = t_i \text{ for some } i, \text{ and } \varphi\{x/t\} \wedge \chi \text{ false}] \end{aligned}$$

This proves  $G(\varphi)$  for  $S(\varphi)$  false.

Case 1 and 2 together prove  $G(\varphi)$  for all  $\varphi$ , hence  $\mu$  is Gaifman. ■

## 4 Probabilities on Interpretations

We now study probabilities defined on sets of interpretations.

Consider the set  $\mathcal{I}$  of interpretations (for the alphabet). A Borel  $\sigma$ -algebra can be defined on  $\mathcal{I}$ . For that, a topology needs to be defined first. Given some alphabet, let  $\mathcal{S}$  denote the set of sentences based on the alphabet. For each sentence  $\varphi$ , let  $\text{mod}(\varphi)$  denote the set

$$\{I \in \mathcal{I} \mid \varphi \text{ is valid in } I\}.$$

Consider the set  $\mathcal{B}_\mathcal{S} = \{\text{mod}(\varphi) \mid \varphi \in \mathcal{S}\}$ . Since  $\mathcal{B}_\mathcal{S}$  is closed under finite intersections, it is a basis for a topology  $\mathcal{T}$  on  $\mathcal{I}$ .  $\mathcal{B}_\mathcal{S}$  is also an algebra, since it is closed under complementation and finite unions, and  $\mathcal{I} \in \mathcal{B}_\mathcal{S}$ . Let  $\mathcal{B}$  be the Borel  $\sigma$ -algebra formed from the topology  $\mathcal{T}$  on  $\mathcal{I}$ . In the following, probabilities on  $\mathcal{B}$  will be considered.

Suppose that the alphabet is countable (equivalently, the set of constants is countable). Then the set of terms and, in particular, the set  $\mathcal{S}$  is countable. In this case,  $\mathcal{B}_\mathcal{S}$  is countable and hence the  $\sigma$ -algebra generated by  $\mathcal{B}_\mathcal{S}$  is the same as the Borel  $\sigma$ -algebra  $\mathcal{B}$  generated by  $\mathcal{T}$ .

**Definition 28 (probability on interpretations)** *A function  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  is a finitely additive probability on algebra  $\mathcal{B}$  if  $\mu^*(\emptyset) = 0$  and  $\mu^*(\mathcal{I}) = 1$  and  $\mu^*(A \cap C) + \mu^*(A \cup C) = \mu^*(A) + \mu^*(C)$  for all  $A, C \in \mathcal{B}$ . It is called a Countably Additive (CA) probability or simply a probability if additionally for all countable collections  $\{A_i\}_{i \in I} \subset \mathcal{B}$  of pairwise disjoint sets with  $\bigcup_{i \in I} A_i \in \mathcal{B}$  it holds that  $\mu^*(\bigcup_{i \in I} A_i) = \sum_{i \in I} \mu^*(A_i)$ .*

For CA-probabilities,  $\mathcal{B}$  is usually assumed to be a Borel  $\sigma$ -algebra, i.e.  $\bigcup_{i \in I} A_i \in \mathcal{B}$  always holds. Countable additivity is equivalent to finite additivity and *continuity*:

$$\lim_{n \rightarrow \infty} \mu^*(\bigcap_{i=1}^n A_i) = \mu^*(\lim_{n \rightarrow \infty} \bigcap_{i=1}^n A_i) \quad \text{for all } A_i \in \mathcal{B}.$$

First we show that a probability on the algebra gives a probability on sentences.

**Proposition 29 ( $\mu^* \Rightarrow \mu$ )** *Let  $\mathcal{S}$  be the set of sentences,  $\mathcal{I}$  the set of interpretations,  $\mathcal{B}_\mathcal{S} = \{\text{mod}(\varphi) \mid \varphi \in \mathcal{S}\}$  the algebra on  $\mathcal{I}$ , and  $\mu^* : \mathcal{B}_\mathcal{S} \rightarrow \mathbb{R}$  a finitely additive probability on  $\mathcal{B}_\mathcal{S}$ . Define  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  by*

$$\mu(\varphi) = \mu^*(\text{mod}(\varphi)),$$

*for each  $\varphi \in \mathcal{S}$ . Then  $\mu$  is a probability on  $\mathcal{S}$ .*

**Proof.** The two conditions of Definition 17 have to be established. Note that  $\mu$  is non-negative because  $\mu^*$  is.

Suppose that  $\varphi$  is valid. Then  $\text{mod}(\varphi) = \mathcal{I}$ , so that  $\mu(\varphi) = \mu^*(\text{mod}(\varphi)) = \mu^*(\mathcal{I}) = 1$ .

Suppose that  $\neg(\varphi \wedge \psi)$  is valid. Hence  $\text{mod}(\varphi) \cap \text{mod}(\psi) = \emptyset$ . Thus

$$\begin{aligned} & \mu(\varphi \vee \psi) \\ &= \mu^*(\text{mod}(\varphi \vee \psi)) \\ &= \mu^*(\text{mod}(\varphi) \cup \text{mod}(\psi)) \\ &= \mu^*(\text{mod}(\varphi)) + \mu^*(\text{mod}(\psi)) \quad [\mu^* \text{ is finitely additive}] \\ &= \mu(\varphi) + \mu(\psi). \end{aligned}$$



Hence  $\mu$  is a probability. ■

Note that only the finite additivity of  $\mu^*$  is needed in Proposition 29.

Next we show that a probability on sentences gives a probability on interpretations. For this, a useful property of probabilities on  $\mathcal{B}_S$  is needed.

**Proposition 30 (finite  $\Leftrightarrow$  countable additivity)** *Let  $\mathcal{S}$  be the set of sentences,  $\mathcal{I}$  the set of interpretations, and  $\mathcal{B}_S = \{mod(\varphi) \mid \varphi \in \mathcal{S}\}$  the algebra on  $\mathcal{I}$ . Then every finitely additive probability on  $\mathcal{B}_S$  is countably additive on  $\mathcal{B}_S$ .*

**Proof.** Let  $\mu^*$  be a finitely additive probability on  $\mathcal{B}_S$ . Suppose that  $\{\varphi_n\}_{n=1}^\infty$  is a sequence of sentences such that  $mod(\varphi_n) \supseteq mod(\varphi_{n+1})$ , for  $n = 1, 2, \dots$ , and  $\bigcap_{n=1}^\infty mod(\varphi_n) = \emptyset$ . Clearly  $\varphi_{n+1} \rightarrow \varphi_n$  is valid, for  $n = 1, 2, \dots$ . Next we claim that  $\varphi_{n_0}$  is unsatisfiable, for some  $n_0$ . To prove this, suppose on the contrary that  $\varphi_n$  is satisfiable, for  $n = 1, 2, \dots$ . Since  $\varphi_{n+1} \rightarrow \varphi_n$  is valid, for  $n = 1, 2, \dots$ , it follows that  $\{\varphi_1, \dots, \varphi_n\}$  is satisfiable, for  $n = 1, 2, \dots$ . By the compactness theorem,  $\{\varphi_n\}_{n=1}^\infty$  is satisfiable, which contradicts the assumption that  $\bigcap_{n=1}^\infty mod(\varphi_n) = \emptyset$ . Thus the claim that  $\varphi_{n_0}$  is unsatisfiable, for some  $n_0$ , is proved. Since the  $mod(\varphi_n)$  are decreasing, we have that  $mod(\varphi_n) = \emptyset$ , for  $n \geq n_0$ . It thus follows that  $\lim_{n \rightarrow \infty} \mu^*(mod(\varphi_n)) = \mu^*(\emptyset) = 0$ . Hence, by [Dud02, Theorem 3.1.1],  $\mu^*$  is countably additive on  $\mathcal{B}_S$ . ■

**Proposition 31 ( $\mu \Rightarrow \mu^*$ )** *Let the alphabet be countable,  $\mathcal{S}$  the set of sentences,  $\mathcal{I}$  the set of interpretations, and  $\mathcal{B}$  the Borel  $\sigma$ -algebra on  $\mathcal{I}$ . Let  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  be a probability on sentences. Then there exists a unique probability  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  such that*

$$\mu^*(mod(\varphi)) = \mu(\varphi),$$

for each  $\varphi \in \mathcal{S}$ .

**Proof.** Consider the algebra  $\mathcal{B}_S = \{mod(\varphi) \mid \varphi \in \mathcal{S}\}$ . Define  $\mu^* : \mathcal{B}_S \rightarrow \mathbb{R}$  by

$$\mu^*(mod(\varphi)) = \mu(\varphi),$$

for each  $\varphi \in \mathcal{S}$ . Suppose that  $\varphi$  and  $\psi$  are sentences such that  $mod(\varphi) = mod(\psi)$ . Then  $\varphi = \psi$  is valid, and so  $\mu(\varphi) = \mu(\psi)$ . This shows that  $\mu^*$  is well-defined on basic sets.

Clearly  $\mu^*(\mathcal{I}) = \mu^*(mod(T)) = \mu(T) = 1$ .

Next it is shown that  $\mu^*$  is finitely additive on the algebra  $\mathcal{B}_S$ . Let  $\{mod(\varphi_i)\}_{i=1}^n$  be a finite collection of pairwise disjoint sets in  $\mathcal{B}_S$ . Suppose that, for some  $i$  and  $j$ ,  $\neg(\varphi_i \wedge \varphi_j)$  is not valid. Hence  $\varphi_i \wedge \varphi_j$  has a model, and so  $mod(\varphi_i) \cap mod(\varphi_j) \neq \emptyset$ . Thus  $mod(\varphi_i) \cap mod(\varphi_j) = \emptyset$  implies  $\neg(\varphi_i \wedge \varphi_j)$  is valid. Then

$$\mu^*\left(\bigcup_{i=1}^n mod(\varphi_i)\right) = \mu^*\left(mod\left(\bigvee_{i=1}^n \varphi_i\right)\right) = \mu\left(\bigvee_{i=1}^n \varphi_i\right) = \sum_{i=1}^n \mu(\varphi_i) = \sum_{i=1}^n \mu^*(mod(\varphi_i)),$$

where the second last equality follows from Part 6 of Proposition 19. Thus  $\mu^*$  is finitely additive on  $\mathcal{B}_S$ .

Now, by Proposition 30,  $\mu^*$  is countably additive on  $\mathcal{B}_S$ . Since the alphabet is countable,  $\mathcal{B}_S$  is countable, and so the Borel  $\sigma$ -algebra  $\mathcal{B}$  generated by the topology on  $\mathcal{I}$  is the same as

the  $\sigma$ -algebra generated by  $\mathcal{B}_S$ . By Caratheodory's theorem [Dud02, Theorem 3.1.4], there is a unique extension of  $\mu^*$  to the Borel  $\sigma$ -algebra  $\mathcal{B}$  on  $\mathcal{I}$ . ■

A probability  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  on sets of interpretations has the following intended meaning:

For a Borel set  $B \in \mathcal{B}$ ,  $\mu^*(B)$  is the degree of belief that the intended interpretation is a member of  $B$ .

We now consider probabilities defined on sets of *separating* interpretations. Let  $\widehat{\mathcal{I}}$  be the set of separating interpretations (for the alphabet). A Borel  $\sigma$ -algebra can be defined on  $\widehat{\mathcal{I}}$ . For that, a topology needs to be defined first. For each sentence  $\varphi$ , let  $\widehat{mod}(\varphi)$  denote the set

$$\{I \in \widehat{\mathcal{I}} \mid \varphi \text{ is valid in } I\}.$$

Consider the set  $\widehat{\mathcal{B}}_S = \{\widehat{mod}(\varphi) \mid \varphi \in \mathcal{S}\}$ . Since  $\widehat{\mathcal{B}}_S$  is closed under finite intersections, it is a basis for a topology  $\widehat{\mathcal{T}}$  on  $\widehat{\mathcal{I}}$ .  $\widehat{\mathcal{B}}_S$  is also an algebra, since it is closed under complementation and finite unions, and  $\widehat{\mathcal{I}} \in \widehat{\mathcal{B}}_S$ . Let  $\widehat{\mathcal{B}}$  be the Borel  $\sigma$ -algebra formed from the topology  $\widehat{\mathcal{T}}$  on  $\widehat{\mathcal{I}}$ . In the following, probabilities on  $\widehat{\mathcal{B}}$  will be considered. The Gaifman condition is crucial for them to be CA, since  $\widehat{\mathcal{B}}$  is not compact unlike  $\mathcal{B}$ .

Suppose that the alphabet is countable. Then the set of terms and, in particular, the set  $\mathcal{S}$  is countable. In this case,  $\widehat{\mathcal{B}}_S$  is countable and hence the  $\sigma$ -algebra generated by  $\widehat{\mathcal{B}}_S$  is the same as the Borel  $\sigma$ -algebra  $\widehat{\mathcal{B}}$  generated by  $\widehat{\mathcal{T}}$ .

Note that there is a one-to-one correspondence between the set of probabilities on  $\widehat{\mathcal{B}}$  and the set of probabilities on  $\mathcal{B}$  which give measure 0 to the set of non-separating interpretations. (The set of non-separating interpretations, and hence the set of separating interpretations, are shown to be  $\mathcal{B}$ -measurable in the proof of Proposition 33 below.) A probability  $\widehat{\mu}^* : \widehat{\mathcal{B}} \rightarrow \mathbb{R}$  can be extended to a probability  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  defined by  $\mu^*(B) = \widehat{\mu}^*(B \cap \widehat{\mathcal{I}})$ , for each  $B \in \mathcal{B}$ . Note that  $\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0$ . Conversely, a probability  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  having the property that  $\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0$  can be restricted to a probability  $\mu^*|_{\widehat{\mathcal{B}}} : \widehat{\mathcal{B}} \rightarrow \mathbb{R}$  defined by  $\mu^*|_{\widehat{\mathcal{B}}}(B) = \mu^*(B)$ , for each  $B \in \widehat{\mathcal{B}}$ .

The next result shows that a probability on the set of separating interpretations gives a Gaifman probability on sentences.

**Proposition 32 (separating  $\mu^* \Rightarrow \mu$  Gaifman)** *Let the alphabet be countable,  $\mathcal{S}$  the set of sentences,  $\widehat{\mathcal{I}}$  the set of separating interpretations, and  $\mu^* : \widehat{\mathcal{B}} \rightarrow \mathbb{R}$  a probability on the Borel  $\sigma$ -algebra  $\widehat{\mathcal{B}}$  on  $\widehat{\mathcal{I}}$ . Define  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  by*

$$\mu(\varphi) = \mu^*(\widehat{mod}(\varphi)),$$

*for each  $\varphi \in \mathcal{S}$ . Then  $\mu$  is a Gaifman probability on  $\mathcal{S}$ .*

**Proof.** First, the two conditions of Definition 17 have to be established. Note that  $\mu$  is non-negative because  $\mu^*$  is.

1. Suppose that  $\varphi$  is valid. Then  $\widehat{mod}(\varphi) = \widehat{\mathcal{I}}$ , so that  $\mu(\varphi) = \mu^*(\widehat{mod}(\varphi)) = \mu^*(\widehat{\mathcal{I}}) = 1$ .
2. Suppose that  $\neg(\varphi \wedge \psi)$  is valid. Hence  $\widehat{mod}(\varphi) \cap \widehat{mod}(\psi) = \emptyset$ . Thus

$$\begin{aligned} & \mu(\varphi \vee \psi) \\ &= \mu^*(\widehat{mod}(\varphi \vee \psi)) \\ &= \mu^*(\widehat{mod}(\varphi) \cup \widehat{mod}(\psi)) \end{aligned}$$

$$\begin{aligned}
&= \mu^*(\widehat{mod}(\varphi)) + \mu^*(\widehat{mod}(\psi)) && [\mu^* \text{ is finitely additive}] \\
&= \mu(\varphi) + \mu(\psi).
\end{aligned}$$

Hence  $\mu$  is a probability.

Let  $r$  and  $s$  be closed terms of type  $\alpha \rightarrow \beta$  and  $t_1, t_2, \dots$  an enumeration of all closed terms of type  $\alpha$ . Then

$$\widehat{mod}(r = s) = \bigcap_{i=1}^{\infty} \widehat{mod}((r \ t_i) = (s \ t_i)).$$

To see this, suppose first that  $I \in \widehat{mod}(r = s)$ . Then clearly  $I \in \widehat{mod}((r \ t_i) = (s \ t_i))$ , for each  $t_i$ . Conversely, suppose that  $I$  is a separating interpretation such that  $I \notin \widehat{mod}(r = s)$ . Since  $I$  is separating, there exists a closed term  $t_j$  such that  $I \notin \widehat{mod}((r \ t_j) = (s \ t_j))$ , for some  $j$ . Hence  $I \notin \bigcap_{i=1}^{\infty} \widehat{mod}((r \ t_i) = (s \ t_i))$ . [Note, by the way, that  $mod(r = s) \neq \bigcap_{i=1}^{\infty} mod((r \ t_i) = (s \ t_i))$ .]

Since  $\forall x.\varphi$  is logically equivalent to  $\lambda x.\varphi = \lambda x.T$ , it follows immediately from the remark of the preceding paragraph that

$$\widehat{mod}(\forall x.\varphi) = \bigcap_{i=1}^{\infty} \widehat{mod}(\varphi\{x/t_i\}).$$

$$\begin{aligned}
\text{Thus } \mu(\forall x.\varphi) &= \mu^*(\widehat{mod}(\forall x.\varphi)) \\
&= \mu^*(\bigcap_{i=1}^{\infty} \widehat{mod}(\varphi\{x/t_i\})) \\
&= \lim_{n \rightarrow \infty} \mu^*(\bigcap_{i=1}^n \widehat{mod}(\varphi\{x/t_i\})) && [\mu^* \text{ is countably additive}] \\
&= \lim_{n \rightarrow \infty} \mu^*(\widehat{mod}(\bigwedge_{i=1}^n \varphi\{x/t_i\})) \\
&= \lim_{n \rightarrow \infty} \mu(\bigwedge_{i=1}^n \varphi\{x/t_i\}),
\end{aligned}$$

and so  $\mu$  is Gaifman, by Proposition 23. ■

A probability  $\mu^* : \widehat{\mathcal{B}} \rightarrow \mathbb{R}$  on sets of separating interpretations has the following intended meaning:

For a Borel set  $B \in \widehat{\mathcal{B}}$ ,  $\mu^*(B)$  is the degree of belief that the intended (separating) interpretation is a member of  $B$ .

Next we show that a Gaifman probability on sentences gives a probability on separating interpretations.

**Proposition 33 (Gaifman  $\mu \Rightarrow \mu^*$  separating)** *Let the alphabet be countable,  $\mathcal{S}$  the set of sentences,  $\widehat{\mathcal{I}}$  the set of separating interpretations, and  $\widehat{\mathcal{B}}$  the Borel  $\sigma$ -algebra on  $\widehat{\mathcal{I}}$ . Let  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  be a Gaifman probability on sentences. Then there exists a unique probability  $\widehat{\mu}^* : \widehat{\mathcal{B}} \rightarrow \mathbb{R}$  such that*

$$\widehat{\mu}^*(\widehat{mod}(\varphi)) = \mu(\varphi),$$

for each  $\varphi \in \mathcal{S}$ .

**Proof.** Let  $\mathcal{I}$  be the set of interpretations. Then  $\mathcal{I} \setminus \widehat{\mathcal{I}}$  is the set of all non-separating interpretations. First we show that  $\mathcal{I} \setminus \widehat{\mathcal{I}}$  is  $\mathcal{B}$ -measurable. Let  $r$  and  $s$  be closed terms of the same function type, say,  $\alpha \rightarrow \beta$ , and  $t_1, t_2, \dots$  an enumeration of all closed terms of type  $\alpha$ . Then

$$\text{mod}(r \neq s) \cap \bigcap_{i=1}^{\infty} \text{mod}((r \ t_i) = (s \ t_i))$$

is a measurable set of non-separating interpretations. Since there are countably many such pairs  $r$  and  $s$ , and since

$$\mathcal{I} \setminus \widehat{\mathcal{I}} = \bigcup_{r,s} \left( \text{mod}(r \neq s) \cap \bigcap_{i=1}^{\infty} \text{mod}((r \ t_i) = (s \ t_i)) \right),$$

it follows immediately that  $\mathcal{I} \setminus \widehat{\mathcal{I}}$  is measurable.

According to Proposition 31, there is a unique probability  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  such that

$$\mu^*(\text{mod}(\varphi)) = \mu(\varphi),$$

for each  $\varphi \in \mathcal{S}$ . We now show that  $\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0$ :

$$\begin{aligned} & \mu^*(\text{mod}(r \neq s) \cap \bigcap_{i=1}^{\infty} \text{mod}((r \ t_i) = (s \ t_i))) \\ &= \mu^*(\bigcap_{i=1}^{\infty} \text{mod}((r \ t_i) = (s \ t_i))) - \mu^*(\text{mod}(r = s)) \\ &= \lim_{n \rightarrow \infty} \mu^*(\bigcap_{i=1}^n \text{mod}((r \ t_i) = (s \ t_i))) - \mu(r = s) \quad [\mu^* \text{ is countably additive}] \\ &= \lim_{n \rightarrow \infty} \mu^*(\text{mod}(\bigwedge_{i=1}^n ((r \ t_i) = (s \ t_i)))) - \mu(r = s) \\ &= \lim_{n \rightarrow \infty} \mu(\bigwedge_{i=1}^n ((r \ t_i) = (s \ t_i))) - \mu(r = s) \\ &= \mu(r = s) - \mu(r = s) \quad [\mu \text{ is Gaifman}] \\ &= 0. \end{aligned}$$

Hence  $\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0$ .

Note that  $\widehat{\mathcal{B}} \subseteq \mathcal{B}$ , since  $\widehat{\mathcal{I}}$  is measurable. Define  $\widehat{\mu}^* : \widehat{\mathcal{B}} \rightarrow \mathbb{R}$  to be the restriction of  $\mu^*$  to  $\widehat{\mathcal{B}}$ . Then, for each  $\varphi \in \mathcal{S}$ ,

$$\begin{aligned} & \widehat{\mu}^*(\widehat{\text{mod}}(\varphi)) \\ &= \mu^*(\text{mod}(\varphi) \cap \widehat{\mathcal{I}}) \\ &= \mu^*(\text{mod}(\varphi)) - \mu^*(\text{mod}(\varphi) \cap (\mathcal{I} \setminus \widehat{\mathcal{I}})) \\ &= \mu^*(\text{mod}(\varphi)) \quad [\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0] \\ &= \mu(\varphi). \end{aligned}$$

Also  $\widehat{\mu}^*(\widehat{\mathcal{I}}) = \mu^*(\widehat{\mathcal{I}}) = \mu^*(\mathcal{I}) - \mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = \mu^*(\mathcal{I}) = 1$ , so that  $\widehat{\mu}^*$  is a probability. ■

Propositions 32 and 33 and imply

**Corollary 34 ( $\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0 \Leftrightarrow \mu$  Gaifman)** *For countable alphabet and any probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  on sentences and probability  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  on interpretations (one-to-one) related by  $\mu^*(\text{mod}(\varphi)) = \mu(\varphi)$  it holds that:  $\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0 \Leftrightarrow \mu$  Gaifman.*

There is a concept of being strongly Cournot for probabilities on sets of interpretations that corresponds to that of being strongly Cournot for probabilities on sentences.

**Definition 35 (strongly Cournot  $\mu^*$ )** *A probability  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  is strongly Cournot if, for each  $\varphi \in \mathcal{S}$ ,  $\varphi$  is satisfiable implies  $\mu^*(\text{mod}(\varphi)) > 0$ .*

**Proposition 36 (strongly Cournot  $\mu^* \Leftrightarrow \mu$ )** *Let  $\mathcal{S}$  be the set of sentences and  $\mathcal{I}$  the set of interpretations. Suppose that  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$ , a probability on the Borel  $\sigma$ -algebra on  $\mathcal{I}$ , and  $\mu : \mathcal{S} \rightarrow \mathbb{R}$ , a probability on sentences, are related by*

$$\mu(\varphi) = \mu^*(\text{mod}(\varphi)),$$

*for each  $\varphi \in \mathcal{S}$ . Then  $\mu$  is a strongly Cournot probability on sentences iff  $\mu^*$  is a strongly Cournot probability on sets of interpretations.*

**Proof.** Suppose that  $\mu$  is a strongly Cournot probability on sentences. Let  $\varphi$  be a satisfiable sentence. Then  $\mu^*(\text{mod}(\varphi)) = \mu(\varphi) > 0$ , and so  $\mu^*$  is a strongly Cournot probability.

Conversely, suppose that  $\mu^*$  is a strongly Cournot probability on sets of interpretations. Let  $\varphi$  be a satisfiable sentence. Then  $\mu(\varphi) = \mu^*(\text{mod}(\varphi)) > 0$ , and so  $\mu$  is a strongly Cournot probability. ■

As with probabilities on sentences, we can also define a Cournot condition for probabilities on sets of separated interpretations.

**Definition 37 (Cournot  $\mu^*$ )** *A probability  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  is Cournot if, for each  $\varphi \in \mathcal{S}$ ,  $\varphi$  has a separating model implies  $\mu^*(\text{mod}(\varphi)) > 0$ .*

Clearly every strongly Cournot probability is Cournot.

**Proposition 38 (Cournot  $\mu^* \Leftrightarrow \mu$ )** *Let  $\mathcal{S}$  be the set of sentences and  $\mathcal{I}$  the set of interpretations. Suppose that  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$ , a probability on the Borel  $\sigma$ -algebra  $\mathcal{B}$  on  $\mathcal{I}$ , and  $\mu : \mathcal{S} \rightarrow \mathbb{R}$ , a probability on sentences, are related by*

$$\mu(\varphi) = \mu^*(\text{mod}(\varphi)),$$

*for each  $\varphi \in \mathcal{S}$ . Then  $\mu$  is a Cournot probability on sentences iff  $\mu^*$  is a Cournot probability on sets of interpretations.*

**Proof.** Suppose that  $\mu$  is a Cournot probability on sentences. Let  $\varphi$  be a sentence having a separating model. Then  $\mu^*(\text{mod}(\varphi)) = \mu(\varphi) > 0$ , and so  $\mu^*$  is a Cournot probability.

Conversely, suppose that  $\mu^*$  is a Cournot probability on sets of interpretations. Let  $\varphi$  be a sentence having a separating model. Then  $\mu(\varphi) = \mu^*(\text{mod}(\varphi)) > 0$ , and so  $\mu$  is a Cournot probability. ■

## 5 Existence of Probabilities

Now we turn to the issue of the existence of probabilities.

**Definition 39 (discrete  $\mu^*$ )** A probability  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  is discrete if there exists a countable set of interpretations  $\{I_i\}_{i=1}^\infty$  and a set of non-negative real numbers  $\{m_i\}_{i=1}^\infty$  such that  $\sum_{i=1}^\infty m_i = 1$  and, for each Borel set  $B$ ,  $\mu^*(B) = \sum_{i: I_i \in B} m_i$ .

Each  $m_i$  is called a *mass*. Clearly, a discrete probability is a probability on the Borel  $\sigma$ -algebra  $\mathcal{B}$ . The set  $\{I_i\}_{i=1}^\infty$  is called the *support* of the probability.

**Theorem 40 (Cournot and Gaifman probability)** *If the alphabet is countable, there exists a probability on sentences that is Cournot and Gaifman.*

**Proof.** Consider an enumeration  $\chi_1, \chi_2, \dots$  of the countable set of sentences which have a separating model. Choose a separating interpretation  $I_i$  in  $\text{mod}(\chi_i)$  and assign the mass  $m_i = \frac{1}{i(i+1)}$  to  $I_i$ , for  $i = 1, 2, \dots$ .

Define  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  to be the discrete probability defined by the masses assigned to this countable set of interpretations. That is, for a Borel set  $B \in \mathcal{B}$ ,  $\mu^*(B) = \sum_{i: I_i \in B} \frac{1}{i(i+1)}$  is the sum of the masses of the subset of separating interpretations in  $\{I_i\}_{i=1}^\infty$  that are members of  $B$ . It is possible that the same interpretation is chosen for more than one  $\text{mod}(\chi_i)$ ; in this case, the masses corresponding to each choice of that interpretation are added together.  $\mu^*$  is a probability, since it is a countable sum of point masses, and  $\mu^*(\mathcal{I}) = \sum_{i=1}^\infty \frac{1}{i(i+1)} = 1$ . Since, for all  $i$ ,  $\mu^*(\text{mod}(\chi_i)) \geq \frac{1}{i(i+1)} > 0$ ,  $\mu^*$  is Cournot.

Now define  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  by  $\mu(\varphi) = \mu^*(\text{mod}(\varphi))$ , for  $\varphi \in \mathcal{S}$ . By Proposition 29,  $\mu$  is a probability on sentences. Also, by Proposition 38,  $\mu$  is Cournot. Finally, note that, if  $\mathcal{I}$  is the set of interpretations and  $\widehat{\mathcal{I}}$  the set of separating interpretations, then  $\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0$ . Consequently, the restriction of  $\mu^*$  to  $\widehat{\mathcal{B}}$  is a probability on  $\widehat{\mathcal{B}}$  and  $\mu(\varphi) = \mu^*(\widehat{\text{mod}}(\varphi))$ , for  $\varphi \in \mathcal{S}$ . Thus, by Proposition 32,  $\mu$  is Gaifman. ■

Note that the support of the discrete probability  $\mu^*$  constructed in Theorem 40 is a dense subset of  $\widehat{\mathcal{I}}$ , since there is a point from the support of the probability in each set in a basis for its topology. Every class of separating models that can be characterized by a finite number of axioms can also be characterized by a single sentence, hence is assigned a non-zero probability.

**Proposition 41 (strongly Cournot probability)** *If the alphabet is countable, there exists a probability on sentences that is strongly Cournot.*

**Proof.** Consider an enumeration  $\chi_1, \chi_2, \dots$  of the countable set of sentences which have a model. Choose an interpretation  $I_i$  in  $\text{mod}(\chi_i)$  and assign the mass  $\frac{1}{i(i+1)}$  to  $I_i$ , for  $i = 1, 2, \dots$ .

Define  $\mu^* : \mathcal{B} \rightarrow \mathbb{R}$  to be the discrete probability defined by  $\mu^*(B) = \sum_{i: I_i \in B} \frac{1}{i(i+1)}$  for  $B \in \mathcal{B}$ .  $\mu^*$  is a probability, since it is a countable sum of point masses, and  $\mu^*(\mathcal{I}) = \sum_{i=1}^\infty \frac{1}{i(i+1)} = 1$ . Since, for all  $i$ ,  $\mu^*(\text{mod}(\chi_i)) \geq \frac{1}{i(i+1)} > 0$ ,  $\mu^*$  is strongly Cournot.

Now define  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  by  $\mu(\varphi) = \mu^*(\text{mod}(\varphi))$ , for  $\varphi \in \mathcal{S}$ . By Proposition 29,  $\mu$  is a probability on sentences. Also, by Proposition 36,  $\mu$  is strongly Cournot. ■

Now we give some illustrative examples concerning the various classes of probabilities that have been introduced.

**Example 42 (a probability which is not Gaifman)** Choose an alphabet for which there exists a non-separating interpretation. Construct  $\mu^*$  by putting unit mass on some non-separating interpretation. The probability on sentences corresponding to  $\mu^*$  is not Gaifman by Corollary 34.

Here is such an alphabet and interpretation. Let there be no non-logical constants in the alphabet. Let the interpretation  $I$  be the standard model defined as follows. The domain  $\mathcal{D}_i = \{d\}$ . Each  $\mathcal{D}_{\alpha \rightarrow \beta}$  consists of all functions from  $\mathcal{D}_\alpha$  to  $\mathcal{D}_\beta$ . Note that  $d$  is not the denotation of any closed term of type  $i$ . Now consider  $\lambda x. \top$  and  $\lambda x. \perp$ , each having type  $i \rightarrow o$ . Clearly  $\mathcal{V}(\lambda x. \top, I) \neq \mathcal{V}(\lambda x. \perp, I)$ . However, there does not exist a closed term  $t$  of type  $i$  such that  $\mathcal{V}((\lambda x. \top) t, I) \neq \mathcal{V}((\lambda x. \perp) t, I)$ . Hence  $I$  is not a separating interpretation.  $\diamond$

Theorem 40 shows that, for any countable alphabet, there is always a probability which is Cournot and Gaifman. The next example shows that it is not guaranteed that there is a probability which is strongly Cournot and Gaifman, because these two concepts may conflict on non-separating interpretations.

**Example 43 (a probability which is strongly Cournot but not Gaifman)** Choose an alphabet for which there exists a non-separating interpretation. Construct  $\mu^*$  by forming an enumeration  $\varphi_1, \varphi_2, \dots$  of all satisfiable sentences, and putting mass  $\frac{1}{2}$  on some non-separating interpretation and for each  $i$  mass  $\frac{1}{(i+1)(i+2)}$  on an interpretation in  $\text{mod}(\varphi_i)$ . The probability on sentences corresponding to  $\mu^*$  is strongly Cournot, but not Gaifman.  $\diamond$

**Example 44 (a probability which is Gaifman but not Cournot)** Choose an alphabet for which there exist two disjoint sentences each having a separating model. Construct  $\mu^*$  by putting unit mass on a separating model of one of the sentences. The probability on sentences corresponding to  $\mu^*$  is Gaifman but not Cournot.

Here is such alphabet and pair of sentences. Let  $d$  be any element,  $\mathcal{D}_i = \{d\}$ , and, for definiteness, each domain  $\mathcal{D}_{\alpha \rightarrow \beta}$  the set of all functions from  $\mathcal{D}_\alpha$  to  $\mathcal{D}_\beta$ . Each of the domains  $\mathcal{D}_\alpha$  is finite. Let there be a non-logical constant  $a$  of type  $i$  such that  $\mathcal{V}(a) = d$ . The domain  $\mathcal{D}_{i \rightarrow o}$  consists of two functions, one that maps  $d$  to  $\top$  and is the denotation of  $\lambda x. \top$ , and one that maps  $d$  to  $\perp$  and is the denotation of  $\lambda x. \perp$ . For each element of each of the domains  $\mathcal{D}_{\alpha \rightarrow \beta}$  (other than  $\mathcal{D}_{i \rightarrow o}$ ) introduce a non-logical constant of a suitable type into the alphabet in such a way that the denotation of the constant is the corresponding function. Note that every element of every domain is the denotation of a closed term. Now introduce a non-logical constant  $p$  of type  $i \rightarrow o$ . For the interpretation  $I_1$ , take everything defined so far and give  $p$  the denotation  $d \mapsto \top$ . Then  $I_1$  is a separating model of the sentence  $(p \ a)$ . On the other hand, for the interpretation  $I_2$  take everything defined so far except give  $p$  the denotation  $d \mapsto \perp$ . Then  $I_2$  is a separating model of the sentence  $\neg(p \ a)$ . Finally, note that  $(p \ a)$  and  $\neg(p \ a)$  are disjoint.

Example 46 below provides another such alphabet and sentence, but with infinite domain  $\mathcal{D}_i = \{0, 1, 2, \dots\}$ : There,  $\forall x. (B \ x)$  and  $\neg \forall x. (B \ x)$  each have a separating model, say  $\hat{I}$  and  $\hat{I}'$ . Hence we can set  $\mu^*(\hat{I}') = 1$ , which implies  $\mu(\forall x. (B \ x)) = 0$  and so  $\mu$  cannot confirm  $\forall x. (B \ x)$ . Note that  $\mu$  is Gaifman by Corollary 34 but not Cournot.  $\diamond$

**Example 45 (a probability which is Cournot but not strongly Cournot)** Choose an alphabet for which there is a sentence having a non-empty set of models all of which are non-separating. Construct  $\mu^*$  by forming an enumeration  $\varphi_1, \varphi_2, \dots$  of all sentences that have a

separating model and putting mass  $\frac{1}{i(i+1)}$  on a separating interpretation in  $\text{mod}(\varphi_i)$ , for each  $i$ . The probability on sentences corresponding to  $\mu^*$  is Cournot but not strongly Cournot.

Here is such an alphabet and sentence. Let the alphabet contain the non-logical constants  $a$  of type  $\iota$  and  $p$  of type  $\iota \rightarrow o$ . Consider the sentence  $\varphi \equiv \exists x.(\neg(p\ x) \wedge (p\ a))$ , which has a model. Let  $I$  be any model for  $\varphi$ . Then for  $I$  the domain  $\mathcal{D}_\iota$  must have at least two elements, one of which is the denotation of  $a$  and where none of the others is the denotation of a closed term of type  $\iota$ . Clearly  $\mathcal{V}(p, I) \neq \mathcal{V}(\lambda x.\top, I)$ . However, there does not exist a closed term  $t$  of type  $\iota$  such that  $\mathcal{V}((p\ t), I) \neq \mathcal{V}((\lambda x.\top\ t), I)$ . Hence  $I$  is not a separating interpretation.  $\diamond$

**Example 46 (standard interpretation of Nat)** This continues Example 24. As non-logical constants in our theory we consider  $\underline{0} : \text{Nat}$  and  $S : \text{Nat} \rightarrow \text{Nat}$ , and abbreviate  $\underline{n} \equiv S^n(\underline{0}) = (S\ (S\ (S\ \dots\ (S\ \underline{0}))))$ . The standard interpretation  $I$  is defined as follows: The domain  $\mathcal{D}_{\text{Nat}} = \{0, 1, 2, \dots\}$ , and each domain  $\mathcal{D}_{\alpha \rightarrow \beta}$  is the set of all functions from  $\mathcal{D}_\alpha$  to  $\mathcal{D}_\beta$ . We interpret  $\mathcal{V}(\underline{n}, I) = n$  and  $V(S) : \mathcal{D}_{\text{Nat}} \rightarrow \mathcal{D}_{\text{Nat}}$  is the successor function mapping  $n$  to  $n + 1$ . This interpretation satisfies the Peano axioms  $\forall x.(S\ x) \neq \underline{0}$  and  $\forall x.\forall y.((S\ x) = (S\ y)) \rightarrow (x = y)$  and  $\forall p.(((p\ \underline{0}) \wedge \forall x.((p\ x) \rightarrow (p\ (S\ x)))) \rightarrow \forall x.(p\ x))$ . We can add to our logic any number of constants of type  $\text{Nat} \rightarrow o$ . Let  $\mathcal{J}$  be the set of interpretations obtained by augmenting  $I$  with any valuation of these new constants. Every interpretation in  $\mathcal{J}$  (still) satisfies the Peano axioms. Here and in later examples we only add one such predicate  $B : \text{Nat} \rightarrow o$ , used for induction. For any probability  $\mu^*$  that concentrates on  $\mathcal{J}$ , i.e.  $\mu^*(\mathcal{J}) = 1$ ,  $\mu(\forall x.(\varphi\ x)) = \lim_{n \rightarrow \infty} \mu((\varphi\ \underline{0}) \wedge \dots \wedge (\varphi\ \underline{n}))$  holds for every closed term  $\varphi$  of type  $\text{Nat} \rightarrow o$ , and in particular for  $B$ .  $\diamond$

**Example 47 (non-standard interpretation of Nat)** Consider Example 46 and modify the interpretation  $I$  to  $I'$  as follows: Expand  $\mathcal{D}_{\text{Nat}}$  to  $\mathcal{D}_{\text{Nat}} = \{0, 1, 2, \dots\} \cup \{\dots, -\tilde{2}, -\tilde{1}, \tilde{0}, \tilde{1}, \tilde{2}, \dots\}$  and  $V(S)$  mapping  $\tilde{n} \mapsto \tilde{n} + 1$  in addition to  $n \mapsto n + 1$ . We call  $\tilde{n} \in \{\dots, -\tilde{2}, -\tilde{1}, \tilde{0}, \tilde{1}, \tilde{2}, \dots\}$ , non-standard numbers. As before, augment  $I'$  by an interpretation of  $B$ . Here we only consider valuations  $V(B)$  that are true everywhere, except on a single non-standard number, say  $\tilde{c}$ . This leads to a non-separating interpretation  $I'$ , since  $\exists x.\neg(B\ x)$  is valid in  $I'$  but there is no closed term  $t$  for which  $\neg(B\ t)$  is. Note that every closed term of type  $\text{Nat}$  has some standard number  $n$  as denotation. For a point probability  $\mu^*$  that concentrates on  $I'$  we therefore have  $\mu(\forall x.(B\ x)) = 0$  but  $\mu((B\ t)) = 1$  for all closed terms  $t$  of type  $\text{Nat}$ . Hence  $\mu$  is not Gaifman and cannot confirm  $\forall x.(B\ x)$ . Note that  $I'$  even satisfies the “Peano” axioms if either  $\forall p$  is replaced by “for all closed terms  $p$  of type  $\text{Nat} \rightarrow o$ ” or a suitable subset of  $\{\top, \text{F}\}^{\mathcal{D}_{\text{Nat}}}$  is chosen for  $\mathcal{D}_{\text{Nat} \rightarrow o}$ . (this is due to the absence of  $+$  and  $\times$ ).  $\diamond$

**Example 48 (the description operator  $\iota$ )** We can use the previous Example 47 to illustrate the complications a description operator  $\iota$  causes. Let constant  $\iota_{(\text{Nat} \rightarrow o) \rightarrow \text{Nat}}$  denote a function that selects the unique member of a singleton set  $((\iota\ (\lambda x.(y = x))) = y)$ . Since  $\mathcal{V}((\iota\ \neg B), I') = \tilde{c}$ ,  $(\iota\ \neg B) = \underline{n}$  is not valid in  $I'$  for any standard number, and  $\mu(B\ (\iota\ \neg B)) = 0$ . Indeed,  $\iota$  makes accessible all non-standard numbers via  $\tilde{c} + k = S^k(\iota\ \neg B)$  and  $\tilde{c} - k = (\iota\ \lambda x.(\neg B\ S^k(x)))$ . Hence  $I'$  is now separating for type  $\text{Nat}$  and all non-standard numbers must be included in the enumeration of terms in the Gaifman condition, even if we only care about the standard interpretation. We do not know how to avoid this problem, e.g. adding additional axioms that constrain  $\iota$ . On the other hand,  $\iota$  can easily be eliminated



from the logic (the basic idea is that formulas like  $(p \text{ } (\iota B))$  can be replaced by something like  $(\exists!x.(B x) \wedge (p x)) \vee (\neg\exists!x.(B x) \wedge (p \underline{0}))$ .  $\diamond$

At least asymptotically, the Cournot and Gaifman probabilities constructed in the proof of Theorem 40 are good priors for sentences, since they are non-dogmatic [GS82]. We will use them in Sections 6 and 7, called  $\xi$  there, to construct minimally more informative distributions given some background knowledge like non-logical axioms.

After having seen various examples of (non)Cournot and (non)Gaifman probabilities, we now give a general characterization of Gaifman and Cournot probabilities.

**Definition 49 (rigid mixture representation)** *Let  $\chi_1, \chi_2, \dots$  be an enumeration of all sentences that have a separating model. We say that a probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  on sentences has a mixture representation iff  $\mu(\varphi) = \sum_{i=1}^{\infty} m_i \mu_i(\varphi)$  for some  $\{m_i > 0\}$  and  $\sum_i m_i = 1$  and probabilities  $\mu_i$  satisfying  $\mu_i(\chi_i) = 1$  (hence  $\mu_i(\neg\chi_i) = 0$ ).*

**Theorem 50 (probability characterization - Gaifman and Cournot)**

*Let  $\mu$  be a probability on sentences. Then*

$$\begin{array}{ll} \mu \text{ is Cournot} & \Leftrightarrow \mu \text{ has a rigid mixture representation} \\ \text{(and Gaifman)} & \text{(and all } \mu_i \text{ in Definition 49 are Gaifman)} \end{array}$$

This result eases the construction of Cournot  $\mu$ , in that it reduces the problem of finding a single  $\mu$  that simultaneously satisfies the infinitely many conditions  $\mu(\chi_i) > 0 \ \forall \chi_i$  to the problem of finding infinitely many probabilities  $\mu_i$  with each only satisfying one constraint  $\mu_i(\chi_i) > 0$ .

For instance, as in the proof of Theorem 40, for any  $I_i \in \widehat{\text{mod}}(\chi_i)$ ,  $\mu_i(\varphi) := \llbracket I_i \in \widehat{\text{mod}}(\varphi) \rrbracket$  satisfies  $\mu_i(\chi_i) = 1$ . This also shows that some Cournot (and Gaifman)  $\mu$  can be built purely from deterministic measures  $\mu_i \in \{0, 1\}$ , i.e. sets of models. Corollary 53 below illustrates more generally how Theorem 50 can help.

**Proof.** With the notation of Definition 49 we have:

(*Cournot* $\Leftarrow$ ) Assume  $\varphi$  has a separating model.

Then  $\varphi = \chi_i$  for some  $i$ , and hence  $\mu(\varphi) = \mu(\chi_i) \geq m_i \mu_i(\chi_i) > 0$ .

(*Gaifman* $\Leftarrow$ ) A linear combination  $\mu$  of Gaifman  $\mu_i$  is itself Gaifman.

(*Cournot* $\Rightarrow$ ) Consider  $\mathbb{N}$ -partition

$\mathcal{T} := \{i \in \mathbb{N} : \mu(\chi_i) = 1\}$ ,

$\mathcal{E} := \{i \notin \mathcal{T} : \chi_i \text{ starts with an even (incl. zero) number of negations } \neg\}$ ,

$\mathcal{O} := \{i \notin \mathcal{T} : \chi_i \text{ starts with an odd number of negations } \neg\}$ .

and let  $c : \mathcal{E} \rightarrow \mathcal{O}$  biject  $\chi_{c(i)} = \neg\chi_i$ . Let  $\varphi$  be an arbitrary sentence.

$$\text{For } i \in \mathcal{E} : \quad \mu(\varphi) = \underbrace{\mu(\varphi|\chi_i)}_{=: \mu_i(\varphi)} \underbrace{\mu(\chi_i)}_{=: p_i > 0} + \underbrace{\mu(\varphi|\neg\chi_i)}_{=: \mu_{c(i)}(\varphi)} \underbrace{\mu(\neg\chi_i)}_{=: 1 - p_i > 0}$$

Let  $\sum_{i \in \mathcal{E}} r_i = 1$  and  $r_i > 0$  and  $m_i = \frac{1}{2} r_i p_i > 0$  and  $m_{c(i)} = \frac{1}{2} r_i (1 - p_i) > 0$  for  $i \in \mathcal{E}$ . Then

$$\mu(\varphi) = \sum_{i \in \mathcal{E}} r_i \mu(\varphi) = \sum_{i \in \mathcal{E}} r_i [p_i \mu_i(\varphi) + (1 - p_i) \mu_{c(i)}(\varphi)] = \sum_{i \in \mathcal{E} \cup \mathcal{O}} 2m_i \mu_i(\varphi)$$

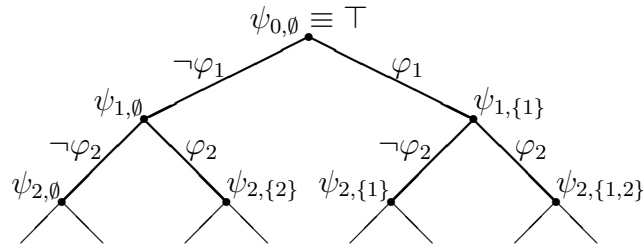
For  $i \in \mathcal{T}$  define  $\mu_i(\varphi) := \mu(\varphi) \equiv \mu(\varphi|\chi_i)$  and  $\sum_{i \in \mathcal{T}} m_i = \frac{1}{2}$  with  $m_i > 0$ . Then  $\mu(\varphi) = \sum_{i \in \mathcal{T}} 2m_i \mu_i(\varphi)$ . Adding both representations gives

$$\mu = \frac{1}{2}[\mu + \mu] = \frac{1}{2} \left[ \sum_{i \in \mathcal{E} \cup \mathcal{O}} 2m_i \mu_i(\varphi) + \sum_{i \in \mathcal{T}} 2m_i \mu_i(\varphi) \right] = \sum_{i=1}^{\infty} m_i \mu_i$$

with  $\sum_{i=1}^{\infty} m_i = 1$ ,  $m_i > 0$ ,  $\mu_i(\chi_i) = 1$  as needed.

( $\mathcal{E}$  Gaifman  $\Rightarrow$ )  $\mu$  Gaifman implies  $\mu_i = \mu(\cdot|\chi_i)$  Gaifman. ■

The next theorem is a complete characterization of general and (strongly) Cournot or Gaifman probabilities on sentences. It is based on a tree construction: Consider a sequence of (some or all) sentences  $\varphi_1, \varphi_2, \varphi_3, \dots$ , arranged in a finite or infinite complete binary tree with all left (right) children at depth  $n$  labeled by  $\neg\varphi_n$  ( $\varphi_n$ ) as depicted below. Furthermore, each node stores the  $\mu$ -probability of the conjunction  $\psi_{n,S}$  of sentences along the edges from the root to this node.



**Proposition 51 ( $\psi_S\varphi$ -tree)** For  $i = 1, \dots, n$ , let  $\varphi_i$  be a sentence. For each  $S \subseteq \{1:n\} \equiv \{1, \dots, n\}$ , define the sentence  $\psi_{n,S}$  by

$$\psi_{n,S} \equiv \left( \bigwedge_{i \in S} \varphi_i \right) \wedge \left( \bigwedge_{j \in \{1:n\} \setminus S} \neg\varphi_j \right).$$

Then the following hold.

1. The  $\psi_{n,S}$ 's are pairwise disjoint.
2.  $\bigvee_{S \subseteq \{1:n\}} \psi_{n,S}$  is valid.
3. For each  $i = 1, \dots, n$ ,  $\varphi_i$  is logically equivalent to  $\bigvee_{S \subseteq \{1:n\}: i \in S} \psi_{n,S}$ .

**Proof.** Straightforward. ■

The following is our main characterization theorem. It states necessary and sufficient conditions on the labels  $\alpha_{n,S} := \mu(\psi_{n,S})$ , for general  $\mu$ , as well as (strongly) Cournot  $\mu$ , and sufficient conditions for Gaifman  $\mu$ . We do not yet have a complete tree characterization of Gaifman probabilities, which is a major open problem. The characterization can easily be converted to a procedure that assigns probabilities to one sentence after the other, but it is not an algorithm, since satisfiability is not decidable.

**Theorem 52 (tree characterization of general/Cournot/Gaifman probabilities)** *Let the alphabet be countable and  $\varphi_1, \varphi_2, \varphi_3, \dots$  an enumeration of all sentences. For each  $n \geq 1$  and each  $S \subseteq \{1:n\}$ , define the sentence  $\psi_{n,S}$  by*

$$\psi_{n,S} \equiv \left( \bigwedge_{i \in S} \varphi_i \right) \wedge \left( \bigwedge_{j \in \{1:n\} \setminus S} \neg \varphi_j \right).$$

1. Let  $\mu$  be a probability on sentences. Then, for each  $n \geq 1$ ,

$$\mu(\varphi_n) = \sum_{S \subseteq \{1:n\}: n \in S} \mu(\psi_{n,S}). \quad (1)$$

Furthermore,  $\mu$  is Cournot (resp., strongly Cournot) iff, for each  $n \geq 1$  and  $S \subseteq \{1:n\}$ ,  $\psi_{n,S}$  has a separating model (resp., is satisfiable) implies  $\mu(\psi_{n,S}) > 0$ .

2. For each  $n \geq 1$  and  $S \subseteq \{1:n\}$ , let  $\alpha_{n,S} \in \mathbb{R}$  satisfy the following conditions.

- (a)  $\alpha_{n,S} \geq 0$ .
- (b) If  $\psi_{n,S}$  is unsatisfiable, then  $\alpha_{n,S} = 0$ .
- (c)  $\alpha_{n,S} = \alpha_{n+1,S} + \alpha_{n+1, S \cup \{n+1\}}$ .
- (d)  $\sum_{S \subseteq \{1:n\}} \alpha_{n,S} = 1$ .

Then there exists a probability  $\mu$  on sentences such that, for each  $n \geq 1$  and each  $S \subseteq \{1:n\}$ ,

$$\mu(\psi_{n,S}) = \alpha_{n,S}.$$

3. Suppose that, in addition to the conditions in Part 2, the following condition also holds: for each  $n \geq 1$  and  $S \subseteq \{1:n\}$ ,  $\psi_{n,S}$  has a separating model (resp., is satisfiable) implies  $\alpha_{n,S} > 0$ . Then  $\mu$  is Cournot (resp., strongly Cournot).
4. Suppose that, the conditions of Part 2 hold. Strengthen 2b by demanding that if  $\psi_{n,S}$  has no separating model, then  $\alpha_{n,S} = 0$ . Further, assume that enumeration  $\varphi_1, \varphi_2, \dots$  is such that if  $\varphi_{n+1} = [r = s]$  for terms  $r$  and  $s$  having the same function type, then  $\varphi_{n+2} = \bigvee_{S \subseteq \{1:n\}} \psi_{n,S} \wedge \varphi\{x/t_S\}$ , where  $\varphi := [(r\ x) = (s\ x)]$  and  $t_S$  is such that  $\psi_{n,S} \wedge \neg \varphi\{x/t_S\}$  has a separating model (if no such  $t_S$  exists, choose  $t_S$  arbitrarily or drop this contribution from  $\bigvee$ ). For  $\varphi_{n+1} = [r = s]$  also set  $\alpha_{n+2,S} = \alpha_{n+1,S}$ . Then  $\mu$  is Gaifman.
5. For every probability  $\mu$ ,  $\alpha_{n,S} := \mu(\psi_{n,S})$  satisfies 2(a)-(d).

Items 1,2,3, and 5 are rather natural. The somewhat ugly item 4 requires explanation: First, the assumption on the enumeration  $\varphi_i$  can easily be satisfied by inserting appropriate  $\varphi_{n+2}$  at the required  $n$ . The intuition behind the construction for  $n = 0$  is that if  $I$  is a model of  $\neg \varphi_1$ , i.e. of  $\exists x. \neg \varphi$ , Gaifman requires a witness  $t$ , which exists by the extensionality axiom. We can guarantee such a witness by putting  $\varphi_2 = \varphi\{x/t\}$  and following exclusively the  $\neg \varphi_2$  branch by setting  $\alpha_{2,\{2\}} = 0$ . For general  $n$ , the witnesses  $t$  and hence  $\varphi_{n+2} = \varphi\{x/t\}$  may depend on  $S$ ; this would lead to a branch-dependent enumeration of sentences. There is nothing wrong with this, and is probably even the preferred solution. In order to keep things simple, we kept the

enumeration branch independent by or-ing  $\varphi_{n+2}$  over all  $2^n$  branches, which makes it formally independent of the branch  $S$ .

**Proof. 1.** The first part follows immediately from Parts 1 and 3 of Proposition 51 and Proposition 19.6.

The second part for strongly Cournot follows immediately from the definition of a strongly Cournot probability, Proposition 51.3, and Proposition 19.4: That strongly Cournot implies  $\mu(\psi_{n,S}) > 0$  for satisfiable  $\psi_{n,S}$  is trivial. For the other direction,  $\varphi_n$  is satisfiable implies that there exists an  $S \ni n$  for which  $\psi_{n,S}$  is satisfiable. Hence  $\mu(\varphi_n) \geq \mu(\psi_{n,S}) > 0$ . Thus  $\mu(\varphi) > 0$ , for all satisfiable  $\varphi$ , and so  $\mu$  is strongly Cournot. The proof for the Cournot case is similar.

**2.** First define  $\mu_0 : \{\psi_{n,S}\}_{n \geq 1, S \subseteq \{1:n\}} \rightarrow \mathbb{R}$  by

$$\mu_0(\psi_{n,S}) = \alpha_{n,S},$$

for each  $n \geq 1$  and  $S \subseteq \{1:n\}$ . We prove by induction that, for  $m \geq n$ ,

$$\mu_0(\psi_{n,S}) = \sum_{R: S \subseteq R \subseteq S \cup \{n+1, \dots, m\}} \alpha_{m,R}.$$

The result is obvious when  $m = n$ . Suppose now it holds for  $m$ . Then

$$\begin{aligned} & \mu_0(\psi_{n,S}) \\ &= \sum_{R: S \subseteq R \subseteq S \cup \{n+1, \dots, m\}} \alpha_{m,R} && \text{[Induction hypothesis]} \\ &= \sum_{R: S \subseteq R \subseteq S \cup \{n+1, \dots, m\}} (\alpha_{m+1,R} + \alpha_{m+1,R \cup \{m+1\}}) \\ &= \sum_{R: S \subseteq R \subseteq S \cup \{n+1, \dots, m+1\}} \alpha_{m+1,R}. \end{aligned}$$

This completes the induction argument.

Now define  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  by

$$\mu(\varphi_n) = \sum_{S \subseteq \{1:n\}: n \in S} \alpha_{n,S}.$$

for each  $n \geq 1$ . We prove by induction that, for  $m \geq n$ ,

$$\mu(\varphi_n) = \sum_{S \subseteq \{1:m\}: n \in S} \alpha_{m,S}.$$

The result is obvious when  $m = n$ . Suppose now it holds for  $m$ . Then

$$\begin{aligned} & \mu(\varphi_n) \\ &= \sum_{S \subseteq \{1:m\}: n \in S} \alpha_{m,S} && \text{[Induction hypothesis]} \\ &= \sum_{S \subseteq \{1:n\}: n \in S} (\alpha_{m+1,S} + \alpha_{m+1,S \cup \{m+1\}}) \\ &= \sum_{S \subseteq \{1:m+1\}: n \in S} \alpha_{m+1,S}. \end{aligned}$$

This completes the induction argument.

We show that  $\mu$  extends  $\mu_0$ . Suppose that  $\psi_{n,S}$ , for some  $n \geq 1$  and  $S \subseteq \{1:n\}$ , is  $\varphi_k$ , for some  $k \geq 1$ . Let  $m = \max\{k, n\}$  and  $\mathcal{A} = \{R : k \in R \subseteq \{1, \dots, m\}\}$  and  $\mathcal{B} = \{R : S \subseteq R \subseteq S \cup \{n+1, \dots, m\}\}$ . Then  $\bigvee_{R \in \mathcal{A}} \psi_{m,R}$  is logically equivalent to  $\varphi_k$  which is equal to  $\psi_{n,S}$

which is logically equivalent to  $\bigvee_{R \in \mathcal{B}} \psi_{m,R}$ . Also the  $\psi_{m,R}$  are pairwise disjoint. Hence  $\psi_{m,R}$  is unsatisfiable (and so  $\alpha_{m,R} = 0$ ) for each  $R \in (\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{B} \setminus \mathcal{A})$ . This implies

$$\mu(\psi_{n,S}) = \mu(\varphi_k) = \sum_{R \in \mathcal{A}} \alpha_{m,R} = \sum_{R \in \mathcal{B}} \alpha_{m,R} = \mu_0(\psi_{n,S}).$$

In summary,  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  is well-defined and satisfies

$$\mu(\varphi_n) = \sum_{S \subseteq \{1:n\}: n \in S} \mu(\psi_{n,S}),$$

for each  $n \geq 1$ .

We show that  $\mu$  is a probability. Clearly,  $\mu$  is non-negative. Now suppose that, for some  $n \geq 1$ ,  $\varphi_n$  is valid.

Then, for  $n \notin S$ ,  $\psi_{n,S}$  is a conjunction that contains  $\neg\varphi_n$ , hence is not satisfiable and therefore  $\alpha_{n,S} = 0$  for  $n \notin S$ . This implies

$$\mu(\varphi_n) = \sum_{S \subseteq \{1:n\}: n \in S} \alpha_{n,S} = \sum_{S \subseteq \{1:n\}} \alpha_{n,S} = 1.$$

Finally, suppose that  $\neg(\varphi_n \wedge \varphi_m)$  is valid. There exists  $k \geq 1$  such  $\varphi_k$  is  $\varphi_n \vee \varphi_m$ . Choose any  $p$  greater than  $n, m$  and  $k$ . Consider  $\mathcal{A} := \{S \subseteq \{1:p\} : k \in S\}$  and  $\mathcal{B} := \{S \subseteq \{1:p\} : n \in S\}$  and  $\mathcal{C} := \{S \subseteq \{1:p\} : m \in S\}$ .

$\alpha_{p,S} = 0$  for  $S \in \mathcal{B} \cap \mathcal{C}$ , since  $\psi_{n,S}$  is a conjunction containing  $\varphi_n \wedge \varphi_m$ .

$\alpha_{p,S} = 0$  for  $\mathcal{A} \setminus (\mathcal{B} \cup \mathcal{C})$ , since  $\psi_{n,S}$  is a conjunction containing  $\varphi_k \wedge \neg\varphi_n \wedge \neg\varphi_m$ .

$\alpha_{p,S} = 0$  for  $(\mathcal{B} \cup \mathcal{C}) \setminus \mathcal{A}$ , since  $\psi_{n,S}$  is a conjunction containing  $\neg\varphi_k \wedge \varphi_n \wedge \varphi_m$ .

Together this implies

$$\mu(\varphi_n \vee \varphi_m) = \mu(\varphi_k) = \sum_{S \in \mathcal{A}} \alpha_{p,S} = \sum_{S \in \mathcal{B}} \alpha_{p,S} + \sum_{S \in \mathcal{C}} \alpha_{p,S} = \mu(\varphi_n) + \mu(\varphi_m).$$

Thus  $\mu$  is a probability on sentences.

**3.** For the strongly Cournot case, suppose that, for some  $n \geq 1$ ,  $\varphi_n$  is satisfiable. Thus  $\psi_{n,S'}$  is satisfiable for some  $S' \subseteq \{1:n\}$  for which  $n \in S'$ . By the condition,  $\mu(\psi_{n,S'}) > 0$ . Hence  $\mu(\varphi_n) = \sum_{S \subseteq \{1:n\}: n \in S} \mu(\psi_{n,S}) > 0$ . The Cournot case is similar.

**4.**  $\exists x.\varphi$  has separating model (s.m.) iff there exists  $t$  such that  $\varphi\{x/t\}$  has s.m. The  $\Rightarrow$  direction follows from Definition 12 with  $r = \lambda x.\varphi$  and  $s = \lambda x.\top$ . The  $\Leftarrow$  direction follows from  $\varphi\{x/t\} \rightarrow \exists x.\varphi$ .

We need to show the Gaifman condition in Definition 20. This is equivalent to: For all terms  $r$  and  $s$  having the same function type,  $\mu(r = s) = \lim_{m \rightarrow \infty} \mu(\bigwedge_{i=1}^m ((r \ t_i) = (s \ t_i)))$ . Fix  $r$  and  $s$ , define  $\varphi := [(r \ x) = (s \ x)]$ . Using the extensionality axiom we hence have to show

$$\mu(\forall x.\varphi) = \lim_{m \rightarrow \infty} \mu\left(\bigwedge_{i=1}^m \varphi\{x/t_i\}\right)$$

Consider  $n$  such that  $\varphi_{n+1} = [r = s] \equiv \forall x.\varphi$ . By assumption,  $\varphi_{n+2} = \bigvee_{S \subseteq \{1:n\}} \psi_{n,S} \wedge \varphi\{x/t_S\}$ .

We first prove that setting  $\alpha_{n+2,S} = \alpha_{n+1,S}$  is allowed:

Assume  $\psi_{n,S} \wedge \neg\varphi_{n+1} \equiv \exists x.(\psi_{n,S} \wedge \neg\varphi)$  has s.m.

$\Rightarrow$  There exists  $t_S$  s.th.  $\psi_{n,S} \wedge \neg\varphi\{x/t_S\}$  has s.m.

$\Rightarrow \psi_{n,S} \wedge \neg\varphi_{n+1} \wedge \neg\varphi\{x/t_S\}$  has s.m., since  $\neg\varphi\{x/t_S\}$  implies  $\neg\varphi_{n+1}$ .

The last expression is logically equivalent to  $\psi_{n,S} \wedge \neg\varphi_{n+1} \wedge \neg\varphi_{n+2}$ , since for  $\psi_{n,S} = \perp$ , both expressions are false, and for  $\psi_{n,S} = \top$ ,  $\psi_{n,S'} = \perp$  for all  $S' \neq S$ , hence  $\bigvee_S$  in  $\varphi_{n+2}$  collapses to  $\varphi\{x/t_S\}$ . Since  $\psi_{n,S} \wedge \neg\varphi_{n+1} \wedge \neg\varphi_{n+2}$  has s.m.,  $\alpha_{n+2,S} = \alpha_{n+1,S}$  is allowed. Assume now that  $\psi_{n,S} \wedge \neg\varphi_{n+1}$  has no s.m. Then  $\psi_{n,S} \wedge \neg\varphi_{n+1} \wedge \neg\varphi_{n+2}$  has neither, and  $\alpha_{n+2,S} = \alpha_{n+1,S} = 0$ . Hence (4.) is a consistent instantiation of (2.) and generates a probability on sentences  $\mu$  with  $\mu(\psi_{n',S'}) = \alpha_{n',S'}$  for all  $n'$  and  $S'$ . We now prove that it is Gaifman.

For  $\mu(\forall x.\varphi \wedge \psi_{n,S}) > 0$ , trivially

$$\mu\left(\bigwedge_{i=1}^m \varphi\{x/t_i\} \mid \forall x.\varphi \wedge \psi_{n,S}\right) = 1 = \mu(\forall x.\varphi \mid \forall x.\varphi \wedge \psi_{n,S})$$

For  $\mu(\neg\forall x.\varphi \wedge \psi_{n,S}) > 0$  and sufficiently large  $m$ ,

$$\mu\left(\bigwedge_{i=1}^m \varphi\{x/t_i\} \mid \neg\forall x.\varphi \wedge \psi_{n,S}\right) = 0 = \mu(\forall x.\varphi \mid \neg\forall x.\varphi \wedge \psi_{n,S})$$

since  $\mu(\neg\varphi\{x/t_S\} \mid \neg\forall x.\varphi \wedge \psi_{n,S}) = \mu(\neg\varphi_{n+2} \mid \neg\forall x.\varphi \wedge \psi_{n,S}) = \alpha_{n+2,S}/\alpha_{n+1,S} = 1$ , and  $\bigwedge_{i=1}^m \varphi\{x/t_i\}$  will eventually contradict  $\neg\varphi\{x/t_S\}$ .

Since both displayed equalities hold for all  $S \subseteq \{1:n\}$ , for sufficiently large  $m$  this implies  $\mu(\bigwedge_{i=1}^m \varphi\{x/t_i\}) = \mu(\forall x.\varphi)$ .

5. Straightforward. ■

Unfortunately items 3 and 4 in Theorem 52 cannot be combined. The  $\mu$  in item 4. is not Cournot, since e.g.  $\neg\varphi_{n+1} \wedge \varphi_{n+2}$  has a separating model if there is more than one possible witness  $t_S$ , but is assigned zero probability. We can do something else though.

The following corollary boosts Gaifman  $\mu$  constructed in Theorem 52.4 with the rigid mixture representation to a Gaifman and Cournot  $\mu$ , and this without having to choose interpretations  $I$  as required in Theorem 50.

**Corollary 53 (Gaifman and Cournot probability)** *Let  $\chi_1, \chi_2, \dots$  be an enumeration of all sentences that have a separating model. For each  $i$ , let  $\varphi_1 := \chi_i, \varphi_2, \varphi_3, \dots$  be different (in the first sentence) enumerations of all sentences, and  $\mu_i$  be a corresponding Gaifman probability constructed in Theorem 52.4, choosing  $\mu_i(\chi_i) \equiv \alpha_{1,\{1\}} := 1$ . Then by Theorem 50, the rigid mixture  $\mu$  of Definition 49 is Gaifman and Cournot.*

## 6 Relative Entropy of Probabilities on Sentences

Assume we “know” the probabilities  $\mu_0(\varphi_i)$  of sentences  $\varphi_1, \dots, \varphi_n$ . Note that  $\mu_0 : \{\varphi_1, \dots, \varphi_n\} \rightarrow [0, 1]$  is *not* a probability on all sentences, but only a partial specification. In the next section (Proposition 57) we derive conditions under which  $\mu_0$  can be extended to a probability over all sentences.

However, if there are any solutions at all, then there are many. It then makes sense to ask whether some distributions that meet our constraints are “better”, in some sense, than others.

A natural idea is to choose  $\mu$  in such a way as to be “as uninformative as possible”, consistent with our constraints as defined by  $\mu_0$ . Unfortunately it is not possible to define “as uninformative as possible” in absolute terms, but we can define it relative to a prior distribution,  $\xi$ . We will formalise this using the concept of relative entropy, or Kullback-Leibler

divergence. We now show that this selection of  $\mu$  has exactly the form of a piecewise re-scaled  $\xi$ , and show how to find the optimal rescaling constants, the various  $\alpha_S$  introduced in the next section, under this criterion. Natural choices for the prior  $\xi$  are the non-dogmatic probabilities constructed in Theorem 40.

We start by introducing the relevant concepts on general measure spaces before constructing the new distribution  $\mu$  that meets our constraints while being uninformative relative to our prior,  $\xi$ .

From [Iha93, p.21][Csi75]: Let  $\mu^*$  and  $\xi^*$  be probabilities on a measurable space  $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ . We say that  $\mu^*$  is *absolutely continuous* with respect to  $\xi^*$ ,  $\mu^* \prec \xi^*$ , if  $\mu^*(A) = 0$  for every  $A \in \mathcal{B}(\mathbf{X})$  such that  $\xi^*(A) = 0$ . By the Radon-Nikodym theorem [Dud02, Theorem 5.5.4], if  $\mu^*$  is absolutely continuous with respect to  $\xi^*$ , then there exists a  $\xi^*$ -integrable function  $\psi(x)$  such that

$$\mu^*(A) = \int_A \psi(x) d\xi^*(x), \quad \forall A \in \mathcal{B}(\mathbf{X}).$$

The function  $\psi(x)$  is called the Radon-Nikodym derivative and is written in the form

$$\psi(x) = \frac{d\mu^*}{d\xi^*}(x).$$

For probabilities  $\mu^*$  and  $\xi^*$  on  $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ , the *relative entropy*  $\text{KL}(\mu^*||\xi^*)$  of  $\mu^*$  with respect to  $\xi^*$  is defined by

$$\text{KL}(\mu^*||\xi^*) := \begin{cases} \int_{\mathbf{X}} \log \frac{d\mu^*}{d\xi^*}(x) d\mu^*(x) & \text{if } \mu^* \prec \xi^*, \\ \infty & \text{otherwise.} \end{cases}$$

The measure  $\xi^*$  is referred to as the *reference measure*.

By reference to this general definition for relative entropy, we can define the relative entropy for probabilities on sentences in two ways:

**Definition 54 (relative entropy on sentences)** *For a countable alphabet and for probabilities  $\mu$  and  $\xi$  defined on some set of sentences  $\mathcal{S}$ , the relative entropy  $\text{KL}(\mu||\xi)$  of  $\mu$  with respect to  $\xi$  is defined by*

$$\text{KL}(\mu||\xi) := \lim_{n \rightarrow \infty} \sum_{S \subseteq \{1:n\}} \mu(\psi_S) \log \frac{\mu(\psi_S)}{\xi(\psi_S)} = \text{KL}(\mu^*||\xi^*)$$

where  $0 \log \frac{0}{\xi} := 0$  and  $\mu \log \frac{\mu}{0} := \infty$  if  $\mu > 0$ . The last equality holds true if  $\mu^*$  and  $\xi^*$  are the probabilities in Proposition 31 on interpretations that correspond to  $\mu$  and  $\xi$  respectively.

The first definition is more general and useful and conceptually easier. Since the relative entropy increases with refinement, the limit always exists and is independent of the order of enumeration of sentences. The second definition is the “obvious” choice for a definition, but is more restrictive and based on much heavier machinery. Equivalence follows from exchanging limits with integrals, which requires some justification.

**Proof.** (sketch) **(i) Order independence:** Let  $\Phi$  be a finite set of sentences, and  $\text{KL}_{\Phi}(\mu||\xi)$  be the relative entropy of the sentences in  $\Phi$ . Then by the monotonicity of the relative entropy under refinement,  $\Phi \subseteq \Phi'$  implies  $\text{KL}_{\Phi} \leq \text{KL}_{\Phi'}$ . It is now routine to establish independence of the limit on the order of enumeration of the sentences.

(ii) *Equivalence of both definitions:* For  $\mu^* \not\prec \xi^*$  one can show that the limit diverges, which implies equality. We will only prove the interesting case when  $\mu^* \prec \xi^*$ . Let  $\varphi_1, \varphi_2, \dots$  be an enumeration of all sentences. For an interpretation  $I \in \mathcal{I}$ , let  $S$  be such that  $\psi_{n,S}$  is valid in  $I$ , i.e.  $S \equiv S(n, I) := \{i \in \{1, \dots, n\} : I \in \text{mod}(\varphi_i)\}$ . Using  $\mu(\psi_{n,S}) = \mu^*(\text{mod}(\psi_{n,S})) = \int_{\text{mod}(\psi_{n,S})} d\mu^*$ , let

$$\begin{aligned} \text{KL}_n(\mu||\xi) &:= \sum_{S \subseteq \{1:n\}} \mu(\psi_{n,S}) \log \frac{\mu(\psi_{n,S})}{\xi(\psi_{n,S})} = \int_{\mathcal{I}} \log \frac{\mu(\psi_{n,S})}{\xi(\psi_{n,S})} d\mu^* \geq 0 \\ \text{KL}^*(\mu||\xi) &:= \int_{\mathcal{I}} \log \frac{d\mu^*}{d\xi^*} d\mu^* \geq 0 \end{aligned}$$

Elementary algebra (telescoping property of KL) allows us to split  $\text{KL}^*$  into a finitary and a tail part

$$\text{KL}^*(\mu||\xi) = \text{KL}_n(\mu||\xi) + \sum_{S \subseteq \{1:n\}} \mu(\psi_{n,S}) \text{KL}^*(\mu(\cdot|\psi_{n,S})||\xi(\cdot|\psi_{n,S}))$$

which shows that  $\text{KL}^* \geq \text{KL}_n$ .

For the other direction, let  $\mathcal{F}_n$  be the Borel  $\sigma$ -algebra generated by  $\{\text{mod}(\psi_{n,S}) : S \subseteq \{1:n\}\}$ . Then  $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$  is a filtration with  $\mathcal{F}_\infty = \mathcal{B}$  the Borel  $\sigma$ -algebra generated by  $\bigcup_{n=1}^\infty \mathcal{F}_n$ . Define

$$Z_n(I) := \frac{\mu^*(\text{mod}(\psi_{n,S}))}{\xi^*(\text{mod}(\psi_{n,S}))} = \frac{\mu(\psi_{n,S})}{\xi(\psi_{n,S})}$$

$Z_n : \mathcal{I} \rightarrow \mathbb{R}$  is an  $\mathcal{F}_n$  measurable function, well-defined with  $\xi$ -probability 1 (w.ξ.p.1).  $Z_1, Z_2, \dots$  forms a  $\xi$ -martingale sequence, since

$$\begin{aligned} \mathbb{E}_\xi[Z_{n+1}|\mathcal{F}_n] &= \frac{\mu(\psi_{n+1,S})}{\xi(\psi_{n+1,S})} \xi(\psi_{n+1,S}|\psi_{n,S}) + \frac{\mu(\psi_{n+1,S \cup \{n+1\}})}{\xi(\psi_{n+1,S \cup \{n+1\}})} \xi(\psi_{n+1,S \cup \{n+1\}}|\psi_{n,S}) \\ &= \frac{\mu(\psi_{n,S})}{\xi(\psi_{n,S})} = Z_n \end{aligned}$$

Since  $\mu^* \prec \xi^*$ , by [Doo53, VII§8] the sequence converges to the Radon-Nikodym derivative

$$\lim_{n \rightarrow \infty} Z_n = \frac{d\mu^*}{d\xi^*} \quad \text{w.}\xi.\text{p.1}$$

Now consider

$$\text{KL}_n(\mu||\xi) = \sum_{S \subseteq \{1:n\}} \frac{\mu(\psi_{n,S})}{\xi(\psi_{n,S})} \log \frac{\mu(\psi_{n,S})}{\xi(\psi_{n,S})} \xi(\psi_{n,S}) = \int_{\mathcal{I}} Z_n \log Z_n d\xi^*$$

By Fatou's lemma applied to  $1 + Z_n \log Z_n$ , which is non-negative, and the existence of the pointwise limit  $Z_n$  w.ξ.p.1, we get

$$\begin{aligned} \liminf_{n \rightarrow \infty} \text{KL}_n(\mu||\xi) &\geq \int_{\mathcal{I}} \liminf_{n \rightarrow \infty} Z_n \log Z_n d\xi^* \\ &= \int_{\mathcal{I}} \frac{d\mu^*}{d\xi^*} \log \frac{d\mu^*}{d\xi^*} d\xi^* = \int_{\mathcal{I}} \log \frac{d\mu^*}{d\xi^*} d\mu^* = \text{KL}^*(\mu||\xi) \end{aligned}$$



Since  $\text{KL}_n$  is monotone increasing and together with  $\text{KL}^* \geq \text{KL}_n$ , we have  $\lim_{n \rightarrow \infty} \text{KL}_n = \text{KL}^*$ . This shows the equivalence of both definitions in Definition 54. ■

Given some base measure  $\xi^*$ , we are interested in finding a measure  $\hat{\mu}^*$  that minimizes  $\text{KL}(\mu^* || \xi^*)$  under some

$$\text{constraints } \int_{\mathbf{X}} f_i(x) d\hat{\mu}^*(x) = a_i, \quad i = 1, \dots, n. \quad (2)$$

We assume that these constraints are satisfiable for some  $\hat{\mu}^* \prec \xi^*$ .

[Iha93] defines the KL-projection of a probability under some constraints as the measure that minimises the relative entropy subject to those constraints. In practice, the KL-projection is defined by giving a Radon-Nikodym derivative that re-scales the original probability to meet the constraints. This is similar to the rescaling used in the proof of Proposition 57 below.

[Iha93, pp.104-5] proves the following: Define functions  $\theta_i(\lambda)$ ,  $i = 1, \dots, n$ , of  $\lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$  by

$$\theta_i(\lambda) = \frac{1}{\Phi(\lambda)} \int_{\mathbf{X}} f_i(x) \exp \left\{ \sum_{j=1}^n \lambda_j f_j(x) \right\} d\xi^*(x), \quad i = 1, \dots, n,$$

$$\text{where } \Phi(\lambda) = \int_{\mathbf{X}} \exp \left\{ \sum_{j=1}^n \lambda_j f_j(x) \right\} d\xi^*(x).$$

We denote by  $\Lambda$  the set of all  $\lambda$  for which the integrals above converge, and define a set  $\mathcal{A} \subseteq (\mathbb{R} \cup \{-\infty\})^n$  by

$$\mathcal{A} = \{(\theta_1(\lambda), \dots, \theta_n(\lambda)); \lambda \in \Lambda\}.$$

Let  $\mathbf{M}_1$  be the set of all probabilities on  $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ ,  $\xi^* \in \mathbf{M}_1$  be a fixed reference measure, and  $f_i(x)$ ,  $i = 1, \dots, n$  be real functions defined on  $\mathbf{X}$ . Assume that  $\mathbf{F} \subset \mathbf{M}_1$  is a set of the form

$$\mathbf{F} = \{\mu^* \in \mathbf{M}_1 : \int_{\mathbf{X}} f_i(x) d\mu^*(x) = a_i, \quad i = 1, \dots, n\},$$

where  $a_i$ ,  $i = 1, \dots, n$  are given constants such that  $(a_1, \dots, a_n) \in \mathcal{A}$ . Then the KL-projection  $\hat{\mu}^*$  on  $\mathbf{F}$  is given by

$$\frac{d\hat{\mu}^*}{d\xi^*}(x) = \frac{1}{\Phi(\lambda)} \exp \left\{ \sum_{i=1}^n \lambda_i f_i(x) \right\}, \quad (3)$$

where  $\lambda = (\lambda_1, \dots, \lambda_n) \in \Lambda$  is a vector uniquely determined by solving

$$\frac{1}{\Phi(\lambda)} \int_{\mathbf{X}} f_i(x) \exp \left\{ \sum_{j=1}^n \lambda_j f_j(x) \right\} d\xi^*(x) = a_i, \quad i = 1, \dots, n.$$

The corresponding minimum relative entropy is given by

$$\text{KL}(\hat{\mu}^* || \xi^*) = \sum_{i=1}^n \lambda_i a_i - \log \Phi(\lambda).$$

We will construct, where possible, a function  $\mu$  that has minimum relative entropy with respect to  $\xi$  while still satisfying our constraints as represented by  $\mu_0$  and the  $\varphi_i$ ,  $i = 1, \dots, n$ . First, we construct a function  $\mu^*$  on interpretations that will end up meeting our constraints while minimising the relative entropy to  $\xi^*$ .

Choose the  $f_i = \llbracket \text{mod}(\varphi_i) \rrbracket$  as indicator function on  $\mathbf{X} = \mathcal{I}$ , which is 1 on models of  $\varphi_i$  and zero elsewhere. Set  $a_i = \mu_0(\varphi_i)$ ,  $i = 1, \dots, n$ . The constraints (2) then reduce to

$$\mu(\varphi_i) = \mu^*(\text{mod}(\varphi_i)) = \int_{\mathcal{I}} \llbracket \text{mod}(\varphi_i) \rrbracket d\mu^* = a_i = \mu_0(\varphi_i)$$

as intended.

Equation (3) then tells us that the scaling function,  $\frac{d\mu^*}{d\xi^*}$ , between  $\xi^*$  and  $\mu^*$  is piecewise constant. In particular,  $\frac{d\mu^*}{d\xi^*}$  is constant across each of the sets  $\text{mod}(\psi_S)$  related to the sentences,  $\psi_S$ , constructed in Proposition 51.

$$\begin{aligned} \mu^*(\text{mod}(\varphi)) &= \int_{\text{mod}(\varphi)} \frac{d\mu^*}{d\xi^*} d\xi^* = \sum_{S \subseteq \{1:n\}} \int_{\text{mod}(\varphi \wedge \psi_S)} \frac{d\mu^*}{d\xi^*} d\xi^* \\ &= \sum_{S \subseteq \{1:n\}} \int_{\text{mod}(\varphi \wedge \psi_S)} \frac{1}{\Phi(\lambda)} \exp \left\{ \sum_{i=1}^n \lambda_i f_i(x) \right\} d\xi^*(x) && [\text{Equation (3)}] \\ &= \sum_{S \subseteq \{1:n\}} \frac{1}{\Phi(\lambda)} \exp \left\{ \sum_{i=1}^n \lambda_i f_i(\text{mod}(\psi_S)) \right\} \xi^*(\text{mod}(\varphi \wedge \psi_S)) && [f_i \text{ constant on } \text{mod}(\psi_S)] \\ &= \frac{1}{\Phi(\lambda)} \sum_{S \subseteq \{1:n\}} \exp \left\{ \sum_{i \in S} \lambda_i \right\} \xi(\varphi \wedge \psi_S) && [f_i = 1 \text{ iff } i \in S] \end{aligned}$$

This leads to the following definition for  $\hat{\mu}$ :

**Definition 55 (minimally more informative probability)** *Let  $\xi$  be an arbitrary probability on sentences, and  $\mu_0 : \{\varphi_1, \dots, \varphi_n\} \rightarrow [0, 1]$  constrain the probability  $\hat{\mu}$  of the sentences  $\varphi_1, \dots, \varphi_n$ . Let*

$$\begin{aligned} \hat{\mu}(\varphi) &:= \sum_{S \subseteq \{1:n\}} w_S \xi(\varphi \wedge \psi_S) && [\text{Defining equation}] \\ w_S &:= \frac{1}{\Phi(\lambda)} \exp \left\{ \sum_{j \in S} \lambda_j \right\} && [\text{Weights}] \\ \Phi(\lambda) &:= \sum_{S \subseteq \{1:n\}} \exp \left\{ \sum_{j \in S} \lambda_j \right\} \xi(\psi_S) && [\text{Normalizing constant}] \\ \mu_0(\varphi_i) &= \sum_{S \subseteq \{1:n\}} w_S \xi(\varphi_i \wedge \psi_S) \equiv \sum_{S \ni i} w_S \xi(\psi_S) && \left[ \begin{array}{l} \text{Consistency equations} \\ \text{for } \lambda_i \in \mathbb{R} \cup \{-\infty\} \end{array} \right] \end{aligned}$$

if the expressions are well-defined and a solution exists. Otherwise  $\hat{\mu}$  is undefined. We call  $\hat{\mu}$  minimally more informative than  $\xi$  given  $\mu_0$  (if it exists).

For  $\varphi = \psi_{S'}$ , only the term  $S = S'$  contributes to the defining equations, which gives the useful relation  $\hat{\mu}(\psi_{S'}) = w_{S'} \xi(\psi_{S'})$ . So indeed,  $w_S = \hat{\mu}(\psi_S)/\xi(\psi_S)$  is the local scaling factor.

Inserting this back into the defining equation, gives

$$\hat{\mu}(\varphi) = \sum_{S:\xi(\psi_S)>0} \hat{\mu}(\psi_S)\xi(\varphi|\psi_S). \quad (4)$$

This also implies that if  $\xi$  is Gaifman, then  $\hat{\mu}(\varphi)$  is Gaifman. Furthermore,  $\hat{\mu}(\varphi) > 0$  whenever consistently with  $\mu_0$  possible and  $\xi(\varphi) > 0$ , i.e. for (strongly) Cournot  $\xi$ ,  $\hat{\mu}$  is as “Cournot” as possible.

**Proposition 56 (minimally more informative probability)** *If  $\mu_0$  can be extended to a probability on  $\mathcal{S}$ , and prior  $\xi(\psi_{n,S}) > 0$  for all satisfiable  $\psi_{n,S}$ , then  $\hat{\mu}$  in Definition 55 is the unique minimum of the relative entropy w.r.t.  $\xi$  under the constraints  $\hat{\mu}(\varphi_i) = \mu_0(\varphi_i)$ ,  $i = 1, \dots, n$ :*

$$\begin{aligned} \min_{\mu:\mu(\varphi_i)=\mu_0(\varphi_i), i=1..n} \{KL(\mu||\xi)\} &= KL(\hat{\mu}||\xi) \\ &= \sum_{S \subseteq \{1:n\}} \hat{\mu}(\psi_S) \log \frac{\hat{\mu}(\psi_S)}{\xi(\psi_S)} = \sum_{i=1}^n \lambda_i \mu_0(\varphi_i) - \log \Phi(\lambda) \end{aligned}$$

**Proof.** A measure-theoretic proof can be based on the second definition in Definition 54 and Equation (3). Here we give an elementary proof based on the first definition: First note that the sum over  $S$  is well defined and finite, since  $\xi(\psi_S) = 0$  implies  $\psi_S$  unsatisfiable implies  $\hat{\mu}(\psi_S) = 0$  by Proposition 19.3. Therefore, wherever necessary or convenient, we interpret sums as being restricted to those  $S$  for which  $\psi_S$  is satisfiable. We have

$$\begin{aligned} KL(\mu||\xi) &= \sum_{S \subseteq \{1:n\}} \mu(\psi_{n,S}) \log \frac{\mu(\psi_{n,S})}{\xi(\psi_{n,S})} \\ &\quad + \lim_{m \rightarrow \infty} \sum_{S \subseteq \{1:n\}} \mu(\psi_{n,S}) \sum_{T \subseteq \{n+1:m\}} \mu(\psi_{m,S \cup T}|\psi_{n,S}) \log \frac{\mu(\psi_{m,S \cup T}|\psi_{n,S})}{\xi(\psi_{m,S \cup T}|\psi_{n,S})} \end{aligned}$$

By multiplying the first term with  $1 = \sum_{T \subseteq \{n+1:m\}} \mu(\psi_{m,S \cup T}|\psi_{n,S})$  and elementary algebra one can easily verify that this expression indeed reduces to the first one in Definition 54. Now we need to minimize this w.r.t. to  $\mu$ . The first term involves a constrained minimization over the  $2^n - 1$  “parameters”  $\mu(\psi_S) : S \subseteq \{1:n\}$ . The second term (for fixed  $m$ ) involves a free minimization over the  $2^n(2^{m-n} - 1)$  parameters  $\mu(\psi_{m,S \cup T}|\psi_{n,S}) : T \subseteq \{n+1:m\}, S \subseteq \{1:n\}$ . Since the two parameter sets are independent, we can minimize both terms separately. Since there are no constraints for the second minimization, and the second term is monotone increasing in  $m$ , the unique solution is obviously  $\mu(\psi_{m,S \cup T}|\psi_{n,S}) = \xi(\psi_{m,S \cup T}|\psi_{n,S})$ . The first term, since  $\xi(\psi_{n,S}) > 0$  and the relative entropy is non-negative and continuous and strictly convex and the domain is finite-dimensional convex and compact (a  $2^n - 1$  dimensional probability simplex), it has a unique minimum on the convex subspace generated by the linear constraints. With Lagrange multipliers and differentiation one can derive the consistency equations in Definition 55, which uniquely determine the solution (this follows the same line of reasoning as after Definition 54, but now in finite sample spaces this is elementary). ■

The next section will develop necessary and sufficient conditions under which  $\mu_0$  can be extended to some  $\mu$  and hence a minimally more informative  $\mu$ .

## 7 Extension of Probabilities

Maintaining consistency in large knowledge bases is a non-trivial problem. Its probabilistic cousin studied in this section is no easier: Given some probabilistic knowledge, does this correspond to a coherent set of probabilistic beliefs?

More formally, suppose a finite set of sentences are given pre-determined probabilities. An interesting, and practically important, question is: what are necessary and sufficient conditions for the existence of a probability on sentences that gives precisely these probabilities on the finite set of sentences? The next result answers this question.

**Proposition 57 (extension of probabilities)** *Let the alphabet be countable alphabet,  $\{\varphi_1, \dots, \varphi_n\}$  be a finite set of sentences, and  $\mu_0 : \{\varphi_1, \dots, \varphi_n\} \rightarrow [0, 1]$  a function. For each  $S \subseteq \{1:n\}$ , let*

$$\psi_S := \left( \bigwedge_{i \in S} \varphi_i \right) \wedge \left( \bigwedge_{j \in \{1:n\} \setminus S} \neg \varphi_j \right).$$

*Then  $\mu_0$  can be extended to a (Gaifman) probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  iff the following set of equations for the  $2^n$  variables  $\alpha_S$ , for  $S \subseteq \{1:n\}$ , has a solution:*

$$\begin{aligned} \sum_{S \subseteq \{1:n\}} \alpha_S &= 1 \\ \sum_{S \subseteq \{1:n\}: i \in S} \alpha_S &= \mu_0(\varphi_i), \text{ for } i = 1, \dots, n \\ \alpha_S &\geq 0, \text{ for } S \subseteq \{1:n\} \\ \alpha_S &= 0 \text{ if } \psi_S \text{ has no (separating) model, for } S \subseteq \{1:n\}. \end{aligned}$$

If the above conditions on  $\alpha_S$  are met, then Proposition 56 and the remark before it imply that  $\mu_0$  can in particular be extended to a probability  $\hat{\mu}$  that is minimally more informative than some prior  $\xi$ , and  $\mu$  is Gaifman if  $\xi$  is.

**Proof.** ( $\Rightarrow$ ) Suppose first that  $\mu_0$  can be extended to a probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$ . We show that the set of equations has a solution.

Define  $\alpha_S = \mu(\psi_S)$ , for each  $S \subseteq \{1:n\}$ . Since the  $\psi_S$ 's are pairwise disjoint, by the definition of a probability, Proposition 19.6, and Proposition 51.2,  $\sum_{S \subseteq \{1:n\}} \alpha_S = 1$ . Also  $\sum_{S \subseteq \{1:n\}: i \in S} \alpha_S = \mu(\varphi_i) = \mu_0(\varphi_i)$ , by Propositions 51.3 and 19.6. Since  $\mu$  is a probability,  $\alpha_S \geq 0$  for  $S \subseteq \{1:n\}$ . Finally,  $\alpha_S = 0$  if  $\psi_S$  is unsatisfiable for  $S \subseteq \{1:n\}$ , by Proposition 19.3. (In case  $\mu$  is Gaifman, we use  $\hat{\mu}^*$  of Proposition 31 to show that  $\alpha_S = \mu(\psi_S) = \hat{\mu}^*(\widehat{mod}(\psi_S)) = \hat{\mu}^*(\emptyset) = 0$  if  $\psi_S$  has no separating model.)

( $\Leftarrow$ ) Conversely, suppose that the equations have a solution. Let  $\xi$  be a strongly Cournot probability on  $\mathcal{S}$  (whose existence is given by Proposition 41). Put

$$Sat = \{S \subseteq \{1:n\} \mid \psi_S \text{ is satisfiable}\}.$$

Define  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  by

$$\mu(\varphi) := \sum_{S \in Sat} \alpha_S \xi(\varphi | \psi_S) = \sum_{S \in Sat} w_S \xi(\varphi \wedge \psi_S) \quad (5)$$

for  $\varphi \in \mathcal{S}$ , where  $w_S := \alpha_S / \xi(\psi_S)$  for  $S \in Sat$ . The function  $\mu$  is well-defined, since  $\xi(\psi_S) > 0$ , if  $\psi_S$  is satisfiable. We claim that  $\mu$  is a probability on sentences. Clearly,  $\mu$  is non-negative.

Suppose that  $\varphi$  is valid. Then

$$\begin{aligned}
& \mu(\varphi) \\
&= \sum_{S \in \text{Sat}} \alpha_S \xi(\varphi | \psi_S) \\
&= \sum_{S \in \text{Sat}} \alpha_S & [\varphi \text{ is valid and } \xi(\cdot | \psi_S) \text{ is a probability}] \\
&= 1. & [\alpha_S = 0 \text{ for } S \notin \text{Sat}]
\end{aligned}$$

Suppose that  $\neg(\varphi \wedge \psi)$  is valid. Then

$$\begin{aligned}
& \mu(\varphi \vee \psi) \\
&= \sum_{S \in \text{Sat}} w_S \xi((\varphi \vee \psi) \wedge \psi_S) & [\text{Equation (5)}] \\
&= \sum_{S \in \text{Sat}} w_S \xi((\varphi \wedge \psi_S) \vee (\psi \wedge \psi_S)) \\
&= \sum_{S \in \text{Sat}} w_S [\xi(\varphi \wedge \psi_S) + \xi(\psi \wedge \psi_S)] & [\neg((\varphi \wedge \psi_S) \wedge (\psi \wedge \psi_S)) \text{ valid}] \\
&= \mu(\varphi) + \mu(\psi).
\end{aligned}$$

Thus  $\mu$  is a probability on sentences.

Finally,  $\mu$  extends  $\mu_0$ :

$$\mu(\varphi_i) = \sum_{S \in \text{Sat}} \alpha_S \xi(\varphi_i | \psi_S) = \sum_{S \in \text{Sat}: i \in S} \alpha_S = \sum_{S \subseteq \{1:n\}: i \in S} \alpha_S = \mu_0(\varphi_i)$$

for  $i = 1, \dots, n$ , which completes the proof. (To proof that  $\mu$  is Gaifman, simply replace ‘is satisfiable’ by ‘has a separating model’ in particular in *Sat*, and ‘ $\xi$  strongly Cournot’ by ‘ $\xi$  Cournot and Gaifman’ in the above proof.) ■

Next we study conditions on the set of sentences which guarantee that the equations of Proposition 57 have a solution. First, a necessary condition is introduced.

**Definition 58 (subadditive  $\mu_0$ )** Let  $\{\varphi_1, \dots, \varphi_n\}$  be a finite set of sentences and  $\mu_0 : \{\varphi_1, \dots, \varphi_n\} \rightarrow [0, 1]$  a function. Then  $\mu_0$  is subadditive if, for each  $i, i_1, \dots, i_k \in \{1, \dots, n\}$  such that the sentences  $\varphi_{i_1}, \dots, \varphi_{i_k}$  are pairwise disjoint and  $\bigvee_{j=1}^k \varphi_{i_j} \rightarrow \varphi_i$  is valid,

$$\begin{aligned}
& \sum_{j=1}^k \mu_0(\varphi_{i_j}) \leq \mu_0(\varphi_i) & \text{and} \\
& \sum_{j=1}^k \mu_0(\varphi_{i_j}) = \mu_0(\varphi_i) & \text{if additionally } \varphi_i \rightarrow \bigvee_{j=1}^k \varphi_{i_j} \text{ is valid.}
\end{aligned}$$

Here is another necessary condition that will be needed.

**Definition 59 (eligible  $\mu_0$ )** Let  $\{\varphi_1, \dots, \varphi_n\}$  be a finite set of sentences and  $\mu_0 : \{\varphi_1, \dots, \varphi_n\} \rightarrow [0, 1]$  a function. Then  $\mu_0$  is eligible if, for each  $i = 1, \dots, n$ ,  $\mu_0(\varphi_i) = 0$  if  $\varphi_i$  is unsatisfiable.

Now the conditions of subadditivity and eligibility are shown to be necessary.

**Proposition 60 (subadditive and eligible  $\mu_0$ )** Let  $\{\varphi_1, \dots, \varphi_n\}$  be a finite set of sentences and  $\mu_0 : \{\varphi_1, \dots, \varphi_n\} \rightarrow [0, 1]$  a function. Suppose that  $\mu_0$  can be extended to a probability on  $\mathcal{S}$ . Then  $\mu_0$  is subadditive and eligible.

**Proof.** Let  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  be a probability that extends  $\mu_0$ .

Suppose that, for some  $i, i_1, \dots, i_k \in \{1, \dots, n\}$ , the sentences  $\varphi_{i_1}, \dots, \varphi_{i_k}$  are pairwise disjoint and  $\bigvee_{j=1}^k \varphi_{i_j} \rightarrow \varphi_i$  is valid. Then

$$\begin{aligned}
& \sum_{j=1}^k \mu_0(\varphi_{i_j}) \\
&= \sum_{j=1}^k \mu(\varphi_{i_j}) && [\mu \text{ extends } \mu_0] \\
&= \mu(\bigvee_{j=1}^k \varphi_{i_j}) && [\text{Proposition 19.6}] \\
&\leq \mu(\varphi_i) && [\text{Proposition 19.4}] \\
&= \mu_0(\varphi_i).
\end{aligned}$$

Also

$$\begin{aligned}
& \varphi_i \rightarrow \bigvee_{j=1}^k \varphi_{i_j} \text{ is valid} \\
&\text{implies } \mu(\bigvee_{j=1}^k \varphi_{i_j}) = \mu(\varphi_i) && [\bigvee_{j=1}^k \varphi_{i_j} \rightarrow \varphi_i \text{ is valid}] \\
&\text{implies } \sum_{j=1}^k \mu(\varphi_{i_j}) = \mu(\varphi_i) \\
&\text{implies } \sum_{j=1}^k \mu_0(\varphi_{i_j}) = \mu_0(\varphi_i).
\end{aligned}$$

Thus  $\mu_0$  is subadditive.

For  $i \in \{1, \dots, n\}$ ,  $\mu(\varphi_i) = 0$  if  $\varphi_i$  is unsatisfiable, since  $\mu$  is a probability; and  $\mu_0(\varphi_i) = \mu(\varphi_i)$ . Hence  $\mu_0$  is eligible.  $\blacksquare$

Now a further structural condition on the set of sentences is introduced that, together with subadditivity and eligibility, will be sufficient to guarantee that there is a solution of the equations.

**Definition 61 (hierarchical sentences)** *A finite set of sentences  $\{\varphi_1, \dots, \varphi_n\}$  is hierarchical if, for each  $i \neq j$ , exactly one of the following holds:  $\neg(\varphi_i \wedge \varphi_j)$  is valid or  $\varphi_i \rightarrow \varphi_j$  is valid or  $\varphi_j \rightarrow \varphi_i$  is valid.*

Intuitively, Definition 61 states that, if  $\varphi_i$  and  $\varphi_j$  ( $i \neq j$ ) are sentences, then either they are disjoint or one of them is stronger than the other. An hierarchical set of sentences is illustrated in Figure 1. Each circle or oval indicates the set of models of a particular sentence.

For the next result, the proof is by induction on the depth of an hierarchical set of sentences; we now define the concept of depth.

**Definition 62 (depth of a sentence)** *Let  $\mathcal{H}$  be an hierarchical set of sentences. The depth of  $\varphi \in \mathcal{H}$  is defined to be the length  $p$  of the unique sequence  $\varphi_1, \dots, \varphi_p \equiv \varphi$  of sentences in  $\mathcal{H}$  such that (a)  $\varphi_{i+1} \rightarrow \varphi_i$  is valid, for  $i = 1, \dots, p-1$ ; (b) for each  $\psi \in \mathcal{H}$ ,  $\varphi_{i+1} \rightarrow \psi$  and  $\psi \rightarrow \varphi_i$  are valid, for some  $i$ , implies  $\psi = \varphi_{i+1}$  or  $\psi = \varphi_i$ ; and (c) for each  $\psi \in \mathcal{H}$ ,  $\varphi_1 \rightarrow \psi$  is valid implies  $\psi = \varphi_1$ .*

*The depth of  $\mathcal{H}$  is the maximum depth of its sentences.*

An empty set of sentences has depth 0. The depth of the set of sentences in Figure 1 is 3.

**Proposition 63 (extending hierarchical constraints)** *Let the alphabet be countable,  $\{\varphi_1, \dots, \varphi_n\}$  a set of sentences, and  $\mu_0 : \{\varphi_1, \dots, \varphi_n\} \rightarrow [0, 1]$  a subadditive eligible function. Suppose that  $\{\varphi_1, \dots, \varphi_n\}$  is hierarchical. Then  $\mu_0$  can be extended to a minimally more informative probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  than some prior  $\xi$  (see Definition 55), which is Gaifman if  $\xi$  is.*

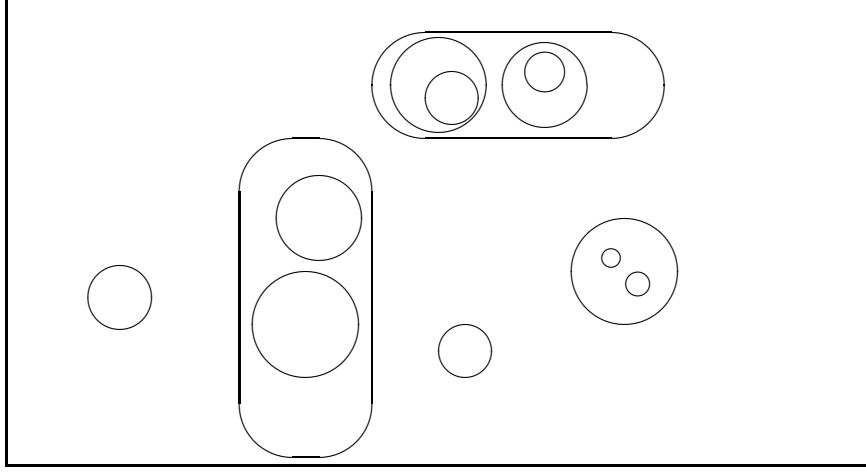


Figure 1: An hierarchical set of sentences

**Proof.** The proof is by induction on the depth  $d$  of the hierarchical set of sentences.

Suppose first that  $d = 0$ , that is, the set of sentences is empty. To show that  $\mu_0$  can be extended to a probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$ , it suffices by Proposition 57 to show that the equations of that proposition for this case have a solution. Since the index set of the set of sentences is empty, its only subset is  $S = \emptyset$ . Furthermore,  $\psi_S$  is  $\top$ . Put  $\alpha_S = 1$ . Then the first equation from Proposition 57 is trivially satisfied. The second set of equations does not appear in this case. Finally, the third and fourth equations are trivially satisfied. This completes the base case of the induction argument.

Now suppose the result holds for hierarchical sets of sentences having depth  $d$ . Let  $\{\varphi_1, \dots, \varphi_n\}$  be an hierarchical set of sentences with depth  $d + 1$ . Without loss of generality, we can assume that  $\{\varphi_1, \dots, \varphi_p\}$ , for  $p < n$ , is an hierarchical set of sentences of depth  $d$  and the sentences  $\varphi_{p+1}, \dots, \varphi_n$  all have depth  $d + 1$ . By the induction hypothesis,  $\mu_0$  restricted to  $\{\varphi_1, \dots, \varphi_p\}$  can be extended to a probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$ . Thus, by Proposition 57, the following set of equations has a solution:

$$\begin{aligned} \sum_{S \subseteq \{1:p\}} \alpha_S &= 1 \\ \sum_{S \subseteq \{1:p\}: i \in S} \alpha_S &= \mu_0(\varphi_i), \text{ for } i = 1, \dots, p \\ \alpha_S &\geq 0, \text{ for } S \subseteq \{1:p\} \\ \alpha_S &= 0 \text{ if } \psi_S \text{ is unsatisfiable, for } S \subseteq \{1:p\}. \end{aligned}$$

Consider a typical sentence  $\varphi_i$  of depth  $d$  that ‘contains’ sentences  $\varphi_{i_1}, \dots, \varphi_{i_k}$  of depth  $d + 1$ . Thus  $\varphi_{i_1}, \dots, \varphi_{i_k}$  are pairwise disjoint and  $\bigvee_{j=1}^k \varphi_{i_j} \rightarrow \varphi_i$  is valid. (See Figure 2.) Since  $\mu_0$  is subadditive,

$$\begin{aligned} \sum_{j=1}^k \mu_0(\varphi_{i_j}) &\leq \mu_0(\varphi_i) && \text{and} \\ \sum_{j=1}^k \mu_0(\varphi_{i_j}) &= \mu_0(\varphi_i) \quad \text{if also} \quad \varphi_i \rightarrow \bigvee_{j=1}^k \varphi_{i_j} \text{ is valid.} \end{aligned}$$

Since  $\mu_0$  is eligible,  $\mu_0(\varphi_{i_j}) = 0$  if  $\varphi_{i_j}$  is unsatisfiable, for  $j = 1, \dots, k$ . It has to be shown that when the depth  $d + 1$  sentences are added to  $\varphi_1, \dots, \varphi_p$ , the corresponding set of equations has a solution.

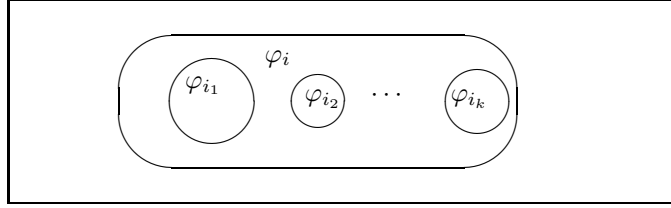


Figure 2: Pairwise disjoint sentences  $\varphi_{i_1}, \dots, \varphi_{i_k}$  of depth  $d + 1$

To simplify the notation, assume for the moment that  $\varphi_{i_1}, \dots, \varphi_{i_k}$  are *all* the sentences of depth  $d + 1$ , so that the index set  $\{1, \dots, p\}$  for the set of at-most-depth  $d$  sentences is expanded to  $\{1, \dots, n\}$  for the whole set of sentences.

Let  $S_i \subseteq \{1, \dots, p\}$  be the set of indices of sentences in the ‘path’ down to  $\varphi_i$  in the set of at-most-depth  $d$  sentences, so that  $\psi_{S_i} = \varphi_i$  is valid. Now consider the full set of sentences. Then the following are valid:

$$\begin{aligned} \psi_{S_i \cup \{i_1\}} &= \varphi_{i_1} \\ &\vdots \\ \psi_{S_i \cup \{i_k\}} &= \varphi_{i_k} \\ \psi_{S_i} &= \varphi_i \wedge \bigwedge_{j=1}^k \neg \varphi_{i_j}. \end{aligned}$$

Included in the equations for the full set of sentences are the following:

$$\begin{aligned} \alpha_{S_i \cup \{i_1\}} &= \mu_0(\varphi_{i_1}) \\ &\vdots \\ \alpha_{S_i \cup \{i_k\}} &= \mu_0(\varphi_{i_k}) \\ \alpha_{S_i} + \alpha_{S_i \cup \{i_1\}} + \dots + \alpha_{S_i \cup \{i_k\}} &= \mu_0(\varphi_i). \end{aligned}$$

(The first  $k$  equations are new ones; the last equation replaces  $\alpha_{S_i} = \mu_0(\varphi_i)$  in the set of equations for the at-most-depth  $d$  sentences.)

Furthermore, the term  $\alpha_{S_i}$  in the first equation of the set of equations for the at-most-depth  $d$  sentences is replaced by  $\alpha_{S_i} + \alpha_{S_i \cup \{i_1\}} + \dots + \alpha_{S_i \cup \{i_k\}}$  in the equations for the full set of sentences. (This is the only change to the first equation because all the other extra subsets  $R$  of  $\{1, \dots, n\}$  that have to be considered lead to  $\psi_R$  that are logically equivalent to  $\perp$  and hence have  $\alpha_R = 0$ .)

Because  $\mu_0$  is subadditive and eligible, it is clear that

$$\begin{aligned} \alpha_S &\geq 0, \text{ for } S \subseteq \{1:n\} \\ \alpha_S &= 0 \text{ if } \psi_S \text{ is unsatisfiable, for } S \subseteq \{1:n\} \end{aligned}$$

are satisfied.

Thus the set of equations for the full set of sentences has a solution. The case when there are extra sentences of depth  $d + 1$  ‘inside’ other  $\varphi_j$  is handled in a similar way.

Now use Propositions 56 and 57 to conclude that  $\mu_0$  can be extended to a minimally more informative probability  $\mu : \mathcal{S} \rightarrow \mathbb{R}$  than some prior  $\xi$ . This completes the induction argument. ■



## 8 User Manual

This section is a brief outlook on how (approximations of) the theory developed in this paper might be used in autonomous reasoning agents. We discuss the special case of certain knowledge and how it can be used to make inferences about statements that are not logical implications of the knowledge base. For instance, if our agent has observed a large number of ravens which are all black without exception, how strongly should it belief in the hypothesis that “all ravens are black”? We construct an agent that can learn in the limit in the usual time-series forecasting setting with an observation sequence indexed by natural numbers.

**Certain knowledge.** A common case of knowledge is a set of sentences  $\varphi_i$ , each having degree of belief 1 (that is,  $\mu_0(\varphi_i) = 1$ , for  $i = 1, \dots, n$ ). In other words, there is certainty that each  $\varphi_i$  is valid in the intended interpretation. This corresponds to non-logical axioms in a theory. Let  $\xi$  be a Cournot probability and suppose that  $\mu$  is minimally more informative than  $\xi$  given  $\mu_0$ . In this case, each  $\mu(\psi_S)$ , for  $S \subseteq \{1:n\}$ , is uniquely determined.

To see this, suppose that  $S \neq \{1:n\}$ , say,  $i \notin S$ . Then  $\mu(\psi_S) \leq \mu(\neg\varphi_i) = 1 - \mu(\varphi_i) = 1 - \mu_0(\varphi_i) = 0$ , so that  $\mu(\psi_S) = 0$ . Hence  $\mu(\psi_{\{1:n\}}) = 1$ . Thus, in this situation, by (4)  $\mu$  satisfies

$$\mu(\varphi) = \xi(\varphi \mid \varphi_1 \wedge \dots \wedge \varphi_n), \quad (6)$$

for  $\varphi \in \mathcal{S}$ . Consequently, there is no optimisation to be done: either  $\varphi_1 \wedge \dots \wedge \varphi_n$  is satisfiable (leading directly to the above definition for  $\mu$ ) or else it is not, in which case there are no solutions and  $\mu$  cannot be defined at all.

A further special case beyond the one just considered is when  $\varphi$  is a logical consequence of  $\varphi_1 \wedge \dots \wedge \varphi_n$ . In this case,

$$\mu(\varphi) = \xi(\varphi \mid \varphi_1 \wedge \dots \wedge \varphi_n) = \frac{\xi(\varphi_1 \wedge \dots \wedge \varphi_n)}{\xi(\varphi_1 \wedge \dots \wedge \varphi_n)} = 1,$$

as one would expect. Similarly when  $\neg\varphi$  is logical consequence, then  $\mu(\varphi) = 0$ .

Note that, while it is important that the prior  $\xi$  be Cournot, it is just as important that the posterior  $\mu$  be allowed not to be Cournot. The prior should be Cournot so that the KL divergence is as widely defined as possible or, more intuitively, to make sure sentences having a separating model are *not forced* to have  $\mu$ -probability 0. On the other hand, the probability  $\mu$  should be *allowed* to be 0 on sentences having a separating model since the evidence in the form of the probabilities on  $\varphi_1, \dots, \varphi_n$  may imply this. This is apparent, for example, for the case where each  $\varphi_i$  has probability 1: according to this evidence, any sentence (even one having a separating model) that is disjoint from  $\varphi_1 \wedge \dots \wedge \varphi_n$  must have  $\mu$ -probability 0.

**Black ravens.** Consider the infamous problem of the black ravens which is one of the most notorious problems in confirmation theory [Ear93, RH11]. Let the ravens be identified by positive integers and  $B(i)$  denote the fact that raven  $i$  is black. The evidence consists of the sentences  $B(1), \dots, B(n)$ . (Thus  $\varphi_i \equiv B(i)$ , for  $i = 1, \dots, n$ .) Let  $\mu_0 : \{B(1), \dots, B(n)\} \rightarrow [0, 1]$  be defined by  $\mu_0(B(i)) = 1$ , for  $i = 1, \dots, n$ . Thus the degree of belief that the  $i$ th raven is black is 1, for  $i = 1, \dots, n$ . Suppose that  $\xi$  is an uninformative prior that is Cournot and Gaifman. Since a-priori there are no constraints (on  $B$ ), this implies that  $\xi(\forall i.B(i)) > 0$ . Let  $\mu$  be a probability that is minimally more informative than  $\xi$  given  $\mu_0$ . Thus  $\mu$  is given by (6).

Now consider the sentence  $\forall i.B(i)$ . This is clearly not a logical consequence of the evidence, but one can use  $\mu$  to ascribe a degree of belief that it is true and, furthermore, investigate what happens to this probability as the number of black ravens increases. Equation (6) and  $\mu_0(B(i)) = 1$ , for  $i = 1, \dots, n$ , and then Theorem 27 applied to Gaifman and Cournot  $\xi$  show that

$$\mu(\forall i.B(i)) = \xi(\forall i.B(i) \mid B(1) \wedge \dots \wedge B(n)) \xrightarrow{n \rightarrow \infty} 1$$

Thus, as the number of observed black ravens increases, the degree of belief that all ravens are black approaches 1. Of course this also implies the weaker statement that our belief in the next raven being black tends to one:

$$\xi(B(n+1) \mid B(1) \wedge \dots \wedge B(n)) \xrightarrow{n \rightarrow \infty} 1$$

**Naive black ravens.** Continuing the preceding example, suppose given the evidence  $B(1), \dots, B(n)$ , each having probability 1, one wants to know the degree of belief for  $B(n+1)$ . Consider the tree construction in Theorem 52 for  $\xi$  but with sentences  $\varphi_1, \varphi_2, \dots$  only ranging over  $\varphi_i = B(i)$  and uniform  $\alpha_{n,S} = 2^{-n}$ . Then

$$\begin{aligned} & \xi(B(n+1) \mid B(1) \wedge \dots \wedge B(n)) \\ &= \frac{\xi(B(1) \wedge \dots \wedge B(n) \wedge B(n+1))}{\xi(B(1) \wedge \dots \wedge B(n))} \\ &= \frac{\alpha_{n+1, \{1:n+1\}}}{\alpha_{n, \{1:n\}}} = 1/2. \end{aligned}$$

Thus, for this prior, knowing the evidence so far, even for large  $n$ , does not give any information about  $B(n+1)$ . But it gets worse: Assume  $\xi$  is somehow extended to a probability on all  $\mathcal{S}$ . Then for any  $m \geq n$ ,

$$\xi(\forall i.B(i) \mid B(1) \wedge \dots \wedge B(n)) \leq \xi(B(1) \wedge \dots \wedge B(m) \mid B(1) \wedge \dots \wedge B(n)) = (\frac{1}{2})^{m-n}$$

hence  $\xi(\forall i.B(i) \mid B(1) \wedge \dots \wedge B(n)) \equiv 0$  for all  $n$ , i.e. universal hypotheses can not be confirmed. Even more seriously, we would be absolutely sure that non-black ravens exist

$$\xi(\exists i.\neg B(i) \mid B(1) \wedge \dots \wedge B(n)) \equiv 1$$

and no number of observed black ravens  $n$  without any counter examples will ever convince us otherwise. These conclusions qualitatively hold even when  $\varphi_1, \varphi_2, \dots$  ranges over all or any subset of quantifier-free/lambda-free sentences. There seem to be no simple local rules for choosing  $\alpha_{n,S}$  that allow confirmation of all universal hypotheses. This shows that it is crucial to include quantified sentences when constructing a prior and ensure it is Cournot (even when only making inferences about unquantified sentences like  $B(n+1)$ ).

**Corollary 64 (learning in the limit)** *Let  $\iota \equiv \text{Nat}$ ,  $\varphi$  be a closed term of type  $\text{Nat} \rightarrow o$ ,  $\mu$  be a Gaifman probability on sentences, and  $\mu(\forall x.(\varphi x)) > 0$ . Then*

$$\lim_{n \rightarrow \infty} \mu(\forall x.(\varphi x) \mid (\varphi \underline{0}) \wedge \dots \wedge (\varphi \underline{n})) = 1$$

This generalizes the black raven example and follows from Theorem 27. In particular, learning in the limit is possible for the Gaifman and Cournot probability constructed in the proof of Theorem 40, provided  $\forall x.(\varphi x)$  has a separating model.

The proof crucially exploits that  $\underline{0}, \underline{1}, \underline{2}, \dots$  are representatives of all terms of type *Nat*. As discussed in Example 48, this would no longer be true had we introduced a description operator into our logic. Corollary 64 would break down and universal hypotheses over the natural numbers could not be inductively confirmed, not even asymptotically.

**Approximations.** The construction of Cournot and Gaifman  $\mu$  in the proof of Theorem 40 required to determine particular separating models for  $\chi_i$  and to determine whether they are also models of other sentences  $\varphi$ . This has been eased by Corollary 53, which only requires determining whether sentences  $\psi_{n,S}$  have (no) separating model. Still this is non-decidable.

Assume we had some calculus to determining whether sentences have (no) separating model. Even an asymptotic or approximate or incomplete calculus may be of use. Fix a sequence on-the-fly of all sentences  $\varphi_2, \varphi_3, \dots$  satisfying Theorem 52.4 (once and for all). Determine the subsequence of all sentences  $\chi_1 = \varphi_{j_1}, \chi_2 = \varphi_{j_2}, \dots$  with separating models (on the fly).

In order to determine  $\mu$  to accuracy  $\varepsilon > 0$  for some finite number of sentences  $\{\varphi_{i_1}, \dots, \varphi_{i_n}\}$  of interest, we have to perform the tree construction “only” for  $\varphi_1 \in \{\chi_1, \dots, \chi_m\}$ , where  $\sum_{i=m+1}^{\infty} < \varepsilon$  and up to depth  $d = \max\{i_1, \dots, i_n\}$ , i.e. determine finitely many cases. If a new sentence  $\varphi_{i_{n+1}}$  of interest “arrives” or higher precision is needed,  $d$  respectively  $m$  can be increased appropriately (that’s what was meant with on-the-fly). It is important to expand the already existing trees with assigned probabilities, rather than restarting the procedure with a larger  $d$ , since this can lead to wrong inductive limits if different choices are made every time.

**Work flow example for a simple inductive reasoning agent.** Below we present an example of a fictitious inductive reasoning agent. It is fictitious, since many operations are incomputable. In practice one needs to employ approximations at various steps. How to do this is an open problem.

1. Assume the agent has been endowed with some background knowledge e.g. about kinetics, colors, biology, birds, etc. Its knowledge is represented in the form of a hierarchical (Definition 61) set of sentences  $\{\varphi_1, \dots, \varphi_n\}$  that hold for sure ( $\mu_0(\varphi_i) = 1$  for some  $i$ ) or with some probability  $0 < \mu_0(\varphi_i) < 1$  for the other  $i$ . Our expressive higher-order logic provides a convenient way of doing so [LN11].

2. Assume  $\mu_0$  is subadditive and eligible (Definitions 58 and 59). This may not be so easy to achieve, and is akin to the general problem of maintaining consistent knowledge bases.

3. Next, use an approximation of a Gaifman and Cournot  $\xi$  prior, e.g. as defined in the proof of Theorems 40 or Theorem 50 or Corollary 53 or and approximation thereof as outlined above. The agent now constructs via Definition 55 the minimally more informative probability  $\mu$ , which exists by Proposition 63 and is Gaifman by Proposition 57 and the remark after Equation (4).

4. Let  $o_0, o_1, o_2, \dots$  be the agent’s life-time sequence of past and future observations of all kinds of objects, ravens and otherwise, all it has/will ever observe, e.g.  $o_n$  is what the agent sees  $n$  seconds after it has been switched on.

5. Assume current time is  $n$ , and the agent needs to hypothesize about the world to decide its next action, e.g. whether some observed regularity is “real”. For instance, “if observation at time  $k$  is a raven, is it also black?”. We can formalize this with a predicate  $\varphi$  of type  $\text{Nat} \rightarrow o$  with the intended interpretation of  $(\varphi \underline{k})$  as “if observation at time  $k$  is a raven, it is black”.

6. Of course the answer to  $(\varphi \underline{0}), \dots, (\varphi \underline{n})$  is immediate, since  $o_0, \dots, o_n$  have already been observed. If they are all true, the agent may start to wonder whether “all ravens are black”, or formally, whether  $\forall x.(\varphi x)$  is true. Note that non-raven observations in the sequence are allowed.

7. If the agent is equipped with our inductive reasoning system, its degree of belief in this hypothesis is  $\mu(\forall x.(\varphi x) | (\varphi \underline{0}) \wedge \dots \wedge (\varphi \underline{n}))$ .

8. This result can be the basis for some decision process maximizing some utilities resulting in an informed action.

Is the degree of belief derived in Step 7 and used in Step 8 reasonable? At least asymptotically Corollary 64 ensures that in the limit the agent’s belief tends to 1, which is very reasonable. So our system of inductive reasoning at least passes this test. Most other inductive reasoning systems have difficulties in getting this right [RH11].

**The Monty Hall Problem.** The Monty Hall problem is based on a US game show. A contestant is presented with three doors. Behind one of the doors is a prize. The other two doors have nothing behind them. The contestant is asked to select a door. After the contestant selects a door, but before that door is opened, the game host selects and opens one of the other two doors. At this point the contestant is again asked to select their preferred door and will win whatever is behind this final selection.

It is expected that the host will not reveal the prize. This constraint means that the host will always open a door to reveal nothing behind it. This limits the contestant’s second choice to either persisting with the door selected originally, or switching to the remaining door. It is a known, if counterintuitive, result that the best strategy for the contestant is to switch doors.

Let  $\iota \equiv Door$ . We introduce the constants  $D_1, D_2, D_3 : Door$  and

$$\begin{aligned} &playerFirstSelection, hostSelection, prizeDoor : Door \rightarrow o \\ &unique : (Door \rightarrow o) \rightarrow o. \end{aligned}$$

As we shall see, the function *unique* is used to capture the constraint on the preceding three predicates that exactly one door makes each of them true. With those, we can now define a set of sentences:

$$\begin{aligned} \varphi_1 &:= (unique = \lambda p. \exists d. ((p d) \wedge \forall x. ((p x) \longrightarrow x = d))) \wedge \\ &\quad (unique playerFirstSelection) \wedge (unique hostSelection) \wedge (unique prizeDoor) \\ \varphi_2 &:= (prizeDoor d_1) \\ \varphi_3 &:= (prizeDoor d_2) \\ \varphi_4 &:= (playerFirstSelection d_1) \\ \varphi_5 &:= (playerFirstSelection d_2) \\ \varphi_6 &:= \forall d. ((hostSelection d) \longrightarrow (\neg(playerFirstSelection d) \wedge \neg(prizeDoor d))) \\ \varphi_7 &:= (hostSelection d_1) \\ \varphi_8 &:= (hostSelection d_2) \\ \varphi_9 &:= \exists d. ((playerFirstSelection d) \wedge (prizeDoor d)) \end{aligned}$$

Selection of the correct prior is very important for this problem. We require that the prior be symmetric in which door makes the *prizeDoor* predicate true, and which door makes the



Boo52] back to Jacob Bernoulli in 1713. An extensive historical account of this thread can be found in [Hai96]; the idea of putting probabilities on sentences goes back to before [Los55] which contains references to even earlier material; the important Gaifman condition appeared in [Gai64] and was further developed in [GS82]; in [SK66] the theory is developed for infinitary logic; overviews of more recent work from a philosophical perspective can be found in [Háj01, Wil02, Wil08b]. The second thread is that of the knowledge representation and reasoning community in artificial intelligence, of which [Nil86, Hal90, FH94, Hal03, SA07] are typical works. The third thread is that of the machine learning community in artificial intelligence, of which [Mug96, DK03, MMR<sup>+</sup>05, RD06, MR07, dSB07, KD07, Pfe07, GMR<sup>+</sup>08] are typical works.

An important and useful technical distinction that can be made between these various approaches is that the combination of logic and probability can be done externally or internally [Wil08b]: in the external view, probabilities are attached to sentences in some logic; in the internal view, sentences incorporate statements about probability. One can even mix the two cases so that probabilities appear both internally and externally. We now examine each of these in turn.

**Probabilities inside sentences.** In the internal view, the uncertainty is modeled inside the sentences of a theory. For this to be possible, we must make a careful choice of logic; in particular, first-order logic (alone) is not expressive enough for this purpose. There has been a tradition of extending first-order logic with probabilistic extensions [Hal03, Háj01, Wil02]. A good alternative approach, studied in [NL09, NLU08], is to simply adopt higher-order logic. The most crucial property of higher-order logic that we exploit is that it admits so-called higher-order functions which take functions as arguments and/or return functions as results. It is this property that allows the modelling of, and reasoning about, probabilistic concepts directly in higher-order theories.

**Probabilities outside sentences.** In contrast to the internal view, almost all other approaches to integrating logic and probability model uncertainty by putting probabilities outside sentences. This natural idea has been taken up by many researchers and has a large body of theoretical support. Here we follow the lead of Gaifman and Snir for first-order logic in [GS82] (that builds on earlier work in [Gai64]). They showed that, under certain conditions, there is a probability on sentences that is strictly positive on consistent sentences (that is, those that have a model). This is an important property of any probability that is intended to be used as a prior in Bayesian inference. An accessible account of this material can be found in [Par94].

Other such systems, and there are now many of these, include Bayesian logic programs [KD07], Markov logic networks [RD06], and stochastic logic programs [Mug96]. While the intention is usually that the probabilities define (or at least constrain) a distribution on the set of interpretations, some systems take other approaches. For example, the probabilities can be used to define a distribution on proofs or a distribution on programs. For a taxonomy of such systems, see [MR07].

**Conclusion.** This paper provides much of the foundation for the design of an integrated probabilistic reasoning system that can handle probabilities both inside and outside sentences. The main challenge for the future lies in the discovery of reasonable approximation schemes for the different currently incomputable aspects of the general theory.

**Acknowledgements.** The research was partly supported by the Australian Research Council Discovery Project DP0877635 “Foundations and Architectures for Agent Systems”. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

## References

- [And02] P.B. Andrews. *An Introduction to Mathematical Logic and Type Theory: To Truth Through Proof*. Kluwer Academic Publishers, second edition, 2002.
- [Boo54] G. Boole. *An Investigation of the Laws of Thought on which are founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberly, 1854.
- [Boo52] G. Boole. *Studies in Logic and Probability*. Watts & Co, 1952.
- [Chu40] A. Church. A formulation of the simple theory of types. *Journal of Symbolic Logic*, 5:56–68, 1940.
- [Cou43] A. A. Cournot. *Exposition de la théorie des chances et des probabilités*. L. Hachette, Paris, 1843.
- [Csi75] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.
- [DK03] L. De Raedt and K. Kersting. Probabilistic logic learning. *SIGKDD Explorations*, 5(1):31–48, 2003.
- [Doo53] J. L. Doob. *Stochastic Processes*. Wiley, New York, 1953.
- [dSB07] R. de Salvo Braz. *Lifted First-Order Probabilistic Inference*. PhD thesis, University of Illinois at Urbana-Champaign, 2007.
- [Dud02] R.M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.
- [Ear93] J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press, Cambridge, MA, 1993.
- [Far08] W.M. Farmer. The seven virtues of simple type theory. *Journal of Applied Logic*, 6(3):267–286, 2008.
- [FH94] R. Fagin and J.Y. Halpern. Reasoning about knowledge and probability. *Journal of the ACM*, 41(2):340–367, 1994.
- [Fin73] T. L. Fine. *Theories of Probability*. Academic Press, New York, 1973.
- [Gai64] H. Gaifman. Concerning measures in first order calculi. *Israel Journal of Mathematics*, 2(1):1–18, 1964.
- [GMR<sup>+</sup>08] N. D. Goodman, V. K. Mansighka, D. Roy, K. Bonawitz, and J. B. Tenenbaum. Church: a language for generative models. In *Uncertainty in Artificial Intelligence*, 2008.
- [GS82] H. Gaifman and M. Snir. Probabilities over rich languages, testing and randomness. *The Journal of Symbolic Logic*, 47(3):495–548, 1982.

- [Hai96] T. Hailperin. *Sentential Probability Logic*. Lehigh University Press, 1996.
- [Háj01] A. Hájek. Probability, logic and probability logic. In L. Goble, editor, *The Blackwell Guide to Philosophical Logic*, chapter 16, pages 362–384. Blackwell, 2001.
- [Hal90] J.Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46(3):311–350, 1990.
- [Hal03] J.Y. Halpern. *Reasoning about Uncertainty*. MIT Press, 2003.
- [Hen50] L. Henkin. Completeness in the theory of types. *Journal of Symbolic Logic*, 15(2):81–91, 1950.
- [Iha93] S. Ihara. *Information theory for continuous systems*. World scientific publishing, 1993.
- [KD07] K. Kersting and L. De Raedt. Bayesian logic programming: Theory and tool. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.
- [Lei94] D. Leivant. Higher-order logic. In D.M. Gabbay, C.J. Hogger, J.A. Robinson, and J. Siekmann, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 2, pages 230–321. Oxford University Press, 1994.
- [Llo03] J.W. Lloyd. *Logic for Learning: Learning Comprehensible Theories from Structured Data*. Cognitive Technologies. Springer, 2003.
- [LN11] J.W. Lloyd and K.S. Ng. Declarative programming for agent applications. *Autonomous Agents and Multi-Agent Systems*, 23(2):224–272, 2011. DOI: 10.1007/s10458-010-9138-1.
- [Los55] J. Los. On the axiomatic treatment of probability. *Colloquium Mathematicum*, 3:125–137, 1955.
- [MMR<sup>+</sup>05] B. Milch, B. Marthi, S. Russell, D. Sontag, D.L. Ong, and A. Kolobov. BLOG: Probabilistic models with unknown objects. In L.P. Kaelbling and A. Saffiotti, editors, *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1352–1359, 2005.
- [MR07] B. Milch and S. Russell. First-order probabilistic languages: Into the unknown. In S. Muggleton, R. Otero, and A. Tamaddoni-Nezhad, editors, *Inductive Logic Programming: 16th International Conference, ILP 2006*, pages 10–24. Springer, LNAI 4455, 2007.
- [Mug96] S. Muggleton. Stochastic logic programs. In L. De Raedt, editor, *Advances in Inductive Logic Programming*, pages 254–264. IOS Press, 1996.
- [Nil86] N.J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1):71–88, 1986.
- [NL09] K.S. Ng and J. W. Lloyd. Probabilistic reasoning in a classical logic. *Journal of Applied Logic*, 7(2):218–238, 2009. DOI:10.1016/j.jal.2007.11.008.
- [NLU08] K.S. Ng, J.W. Lloyd, and W.T.B. Uther. Probabilistic modelling, inference and learning using logical theories. *Annals of Mathematics and Artificial Intelligence*, 54:159–205, 2008. DOI:10.1007/s10472-009-9136-7.
- [Par94] J.B. Paris. *The Uncertain Reasoner’s Companion*, volume 39 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1994.



- [Pfe07] A. Pfeffer. The design and implementation of IBAL: A general-purpose probabilistic language. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning*, chapter 14. MIT Press, 2007.
- [RD06] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.
- [RH11] S. Rathmanner and M. Hutter. A philosophical treatise of universal induction. *Entropy*, 13(6):1076–1136, 2011.
- [SA07] A. Shirazi and E. Amir. Probabilistic modal logic. In R.C. Holte and A. Howe, editors, *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 489–495, 2007.
- [Sha01] S. Shapiro. Classical logic ii – higher-order logic. In L. Goble, editor, *The Blackwell Guide to Philosophical Logic*, pages 33–54. Blackwell, 2001.
- [Sha06] G. Shafer. Why did Cournot’s principle disappear?, 19 May 2006. Presentation. Ecole des Hautes Etudes en Sciences Sociales, Paris. Slides, URL: <http://www.glennshafer.com/assets/downloads/disappear.pdf>.
- [SK66] D. Scott and P. Krauss. Assigning probabilities to logical formula. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*, pages 219–264. North-Holland, 1966.
- [vBD83] J. van Benthem and K. Doets. Higher-order logic. In D.M. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic*, volume 1, pages 275–330. Reidel, 1983.
- [Wil02] J. Williamson. Probability logic. In D. Gabbay, R. Johnson, H.J. Ohlbach, and J. Woods, editors, *Handbook of the Logic of Inference and Argument: The Turn Toward the Practical*, volume 1 of *Studies in Logic and Practical Reasoning*, pages 397–424. Elsevier, 2002.
- [Wil08a] J. Williamson. Objective bayesian probabilistic logic. *Journal of Algorithms*, 63(4):167–183, 2008.
- [Wil08b] J. Williamson. Philosophies of probability. In A. Irvine, editor, *Handbook of the Philosophy of Mathematics, Volume 4 of the Handbook of the Philosophy of Science*. Elsevier, 2008. In press.

## A List of Notation

$x, y, z$	variables
$t, r, s$	terms
$\alpha, \beta$	type of a term
$o$	type of the booleans
$i$	type of individuals
$\top$	Truth
$\perp$	Falsity
$\varphi, \chi, \psi$	formula = term of type $o$ , called sentence if closed
$\mathcal{S}$	set of all sentences
$\mathcal{I}$	set of interpretations
$\widehat{\mathcal{I}}$	set of separating interpretations
$I$	interpretation
$mod(\varphi)$	$\{I \in \mathcal{I}   \varphi \text{ is valid in } I\} = \text{set of models of } \varphi$
$\widehat{mod}(\varphi)$	$\{I \in \widehat{\mathcal{I}}   \varphi \text{ is valid in } I\} = \text{set of separating models of } \varphi$
$\mathcal{B}$	Borel $\sigma$ -algebra generated by $\{mod(\varphi)   \varphi \in \mathcal{S}\}$ , if alphabet countable
$\widehat{\mathcal{B}}$	Borel $\sigma$ -algebra generated by $\{\widehat{mod}(\varphi)   \varphi \in \mathcal{S}\}$ , if alphabet countable
$\mu, (\hat{\mu})$	(estimated) probability on sentences
$\mu^*, (\hat{\mu}^*)$	probability on sets of (separating) interpretations
$i, j, k, n$	natural numbers used for indexing
$\varphi_1, \varphi_2, \dots$	enumeration of some or all sentences
$S$	$\subseteq \{1:n\} \equiv \{1, \dots, n\} = \text{index of "positive" } \varphi \text{ in } \dots$
$\psi_S \equiv \psi_{n,S}$	$(\bigwedge_{i \in S} \varphi_i) \wedge (\bigwedge_{j \in \{1:n\} \setminus S} \neg \varphi_j) = \text{hierarchical basis}$
$\alpha_{n,S}$	$\mu(\psi_{n,S}) = \text{base probabilities}$
$\xi$	prior probability (usually Gaifman and Cournot)
$\mu_0(\varphi_i)$	$: \{\varphi_1, \dots, \varphi_n\} \rightarrow [0, 1] = \text{constraints on } \mu: \mu(\varphi_i) = \mu_0(\varphi_i)$

## B List of Definitions, Theorems, Examples, ...

Definition 1	type $\alpha$ . . . . .	5
theorem.1		
Definition 2	term $t$ . . . . .	5
theorem.2 theorem.3 theorem.4		
Definition 5	frame $\{\mathcal{D}_\alpha\}_\alpha$ . . . . .	6
theorem.5		
Definition 6	valuation $V$ . . . . .	6
theorem.6		
Definition 7	variable assignment $\nu$ . . . . .	6
theorem.7		
Definition 8	interpretation $\langle \{\mathcal{D}_\alpha\}_\alpha, V \rangle$ . . . . .	6

theorem.8		
Definition 9	satisfiable . . . . .	7
theorem.9		
Definition 10	consistency . . . . .	7
theorem.10		
Definition 11	logical consequence . . . . .	7
theorem.11		
Definition 12	separating interpretation/model . . . . .	7
theorem.12		
Definition 13	extensionally complete . . . . .	7
theorem.13		
Proposition 14	extensionally complete $\Rightarrow$ separating . . . . .	8
theorem.14		
Proposition 15	existence of separating models . . . . .	8
theorem.15		
Theorem 16	compactness . . . . .	8
theorem.16		
Definition 17	probability on sentences . . . . .	9
theorem.17		
Definition 18	pairwise disjoint sentences . . . . .	9
theorem.18		
Proposition 19	properties of probability on sentences . . . . .	9
theorem.19		
Definition 20	Gaifman probability . . . . .	10
theorem.20		
Proposition 21	Gaifman probability . . . . .	11
theorem.21		
Proposition 22	limits for countable alphabet . . . . .	12
theorem.22		
Proposition 23	Gaifman for countable alphabet . . . . .	13
theorem.23		
Example 24	natural numbers <i>Nat</i> . . . . .	14
theorem.24		
Definition 25	strongly Cournot probability . . . . .	14
theorem.25		
Definition 26	Cournot probability . . . . .	14
theorem.26		
Theorem 27	confirming universal hypotheses . . . . .	14
theorem.27		
Definition 28	probability on interpretations . . . . .	16
theorem.28		
Proposition 29	$\mu^* \Rightarrow \mu$ . . . . .	16
theorem.29		
Proposition 30	finite $\Leftrightarrow$ countable additivity . . . . .	17
theorem.30		
Proposition 31	$\mu \Rightarrow \mu^*$ . . . . .	17
theorem.31		

Proposition 32	separating $\mu^* \Rightarrow \mu$ Gaifman	18
theorem.32		
Proposition 33	Gaifman $\mu \Rightarrow \mu^*$ separating	19
theorem.33		
Corollary 34	$\mu^*(\mathcal{I} \setminus \widehat{\mathcal{I}}) = 0 \Leftrightarrow \mu$ Gaifman	20
theorem.34		
Definition 35	strongly Cournot $\mu^*$	21
theorem.35		
Proposition 36	strongly Cournot $\mu^* \Leftrightarrow \mu$	21
theorem.36		
Definition 37	Cournot $\mu^*$	21
theorem.37		
Proposition 38	Cournot $\mu^* \Leftrightarrow \mu$	21
theorem.38		
Definition 39	discrete $\mu^*$	22
theorem.39		
Theorem 40	Cournot and Gaifman probability	22
theorem.40		
Proposition 41	strongly Cournot probability	22
theorem.41		
Example 42	a probability which is not Gaifman	23
theorem.42		
Example 43	a probability which is strongly Cournot but not Gaifman	23
theorem.43		
Example 44	a probability which is Gaifman but not Cournot	23
theorem.44		
Example 45	a probability which is Cournot but not strongly Cournot	23
theorem.45		
Example 46	standard interpretation of <i>Nat</i>	24
theorem.46		
Example 47	non-standard interpretation of <i>Nat</i>	24
theorem.47		
Example 48	the description operator $\iota$	24
theorem.48		
Definition 49	rigid mixture representation	25
theorem.49		
Theorem 50	probability characterization - Gaifman and Cournot	25
theorem.50		
Proposition 51	$\psi_S \varphi$ -tree	26
theorem.51		
Theorem 52	tree characterization of general/Cournot/Gaifman probabilities	27
theorem.52		
Corollary 53	Gaifman and Cournot probability	30
theorem.53		
Definition 54	relative entropy on sentences	31
theorem.54		
Definition 55	minimally more informative probability	34

theorem.55		
Proposition 56	minimally more informative probability . . . . .	35
theorem.56		
Proposition 57	extension of probabilities . . . . .	36
theorem.57		
Definition 58	subadditive $\mu_0$ . . . . .	37
theorem.58		
Definition 59	eligible $\mu_0$ . . . . .	37
theorem.59		
Proposition 60	subadditive and eligible $\mu_0$ . . . . .	37
theorem.60		
Definition 61	hierarchical sentences . . . . .	38
theorem.61		
Definition 62	depth of a sentence . . . . .	38
theorem.62		
Proposition 63	extending hierarchical constraints . . . . .	38
theorem.63		
Corollary 64	learning in the limit . . . . .	42
theorem.64		